This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

09/472,249

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: １９９９年１２月 ２日

出 願 番 号
Application Number: 平成１１年特許願第３４３８９０号

出 願 人
Applicant (s): 株式会社リコー

２０００年 １月 ７日

特 許 庁 長 官
Commissioner,
Patent Office 近 藤 隆 彦

出証番号 出証特平１１－３０９１７２６

【書類名】　特許願

【整理番号】　9907814

【提出日】　平成11年12月 2日

【あて先】　特許庁長官殿

【国際特許分類】　G06F 15/40

G06F 17/27

G06F 3/00

【発明の名称】　文書処理装置、文書分類装置、文書処理方法、文書分類方法およびそれらの方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【請求項の数】　38

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　嶋田　敦夫

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　宮地　達生

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　剣持　栄治

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　山崎　真湖人

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　武谷　一寿

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

出証特平１１－３０９１７２６

【氏名】　長束　哲郎

【特許出願人】

【識別番号】　000006747

【氏名又は名称】　株式会社リコー

【代理人】

【識別番号】　100104190

【弁理士】

【氏名又は名称】　酒井　昭徳

【先の出願に基づく優先権主張】

【出願番号】　平成10年特許願第376576号

【出願日】　平成10年12月24日

【先の出願に基づく優先権主張】

【出願番号】　平成10年特許願第369589号

【出願日】　平成10年12月25日

【先の出願に基づく優先権主張】

【出願番号】　平成11年特許願第 22915号

【出願日】　平成11年 1月29日

【手数料の表示】

【予納台帳番号】　041759

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9810808

【プルーフの要否】　要

【書類名】　　明細書

【発明の名称】　文書処理装置、文書分類装置、文書処理方法、文書分類方法およびそれらの方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【特許請求の範囲】

【請求項１】　入力された複数の文書データを所定の形式で表示または印刷するために出力する文書処理装置において、

入力された文書データを記憶する文書記憶手段と、

前記文書記憶手段により記憶された文書データの全部または一部を選択する選択手段と、

前記選択手段により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出手段と、

前記特徴抽出手段により抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理手段と、

前記加工処理手段により加工処理された文書データの全部または一部を出力する出力手段と、

を備えたことを特徴とする文書処理装置。

【請求項２】　前記出力手段は、

前記加工処理手段により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定手段と、

前記項目値設定手段により設定された項目値ごとに前記文書データの全部または一部を集計する集計手段と、

を備え、

前記文書データの全部または一部を、項目値を少なくとも一つの軸とする表形式に展開して出力することを特徴とする請求項１に記載の文書処理装置。

【請求項３】　前記出力手段は、さらに、前記加工処理手段により加工処理された文書データの全部または一部を、前記加工処理手段により加工処理される前の文書データの全部または一部とともに出力することを特徴とする請求項１または２に記載の文書処理装置。

【請求項４】　前記文書記憶手段は、さらに、前記加工処理手段により加工処理された文書データの全部または一部を記憶することを特徴とする請求項１～３のいずれか一つに記載の文書処理装置。

【請求項５】　前記選択手段は、さらに、前記出力手段により出力された文書データの全部または一部を選択することを特徴とする請求項１～４のいずれか一つに記載の文書処理装置。

【請求項６】　前記文書記憶手段は、さらに、前記加工処理の内容に関するデータを記憶することを特徴とする請求項１～５のいずれか一つに記載の文書処理装置。

【請求項７】　文書の内容に基づいて文書の分類をおこなう文書分類装置において、

文書データを入力する入力手段と、

前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、

前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、

前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、

前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、

前記クラスタ特徴算出手段により算出されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶手段と、

を備えたことを特徴とする文書分類装置。

【請求項８】　文書の内容に基づいて文書の分類をおこなう文書分類装置において、

文書データを入力する入力手段と、

前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、

前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対

する文書特徴ベクトルを生成するベクトル生成手段と、

前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、

前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、

前記クラスタ特徴算出手段により算出されたクラスタ特徴を表示する表示手段と、

前記分類手段により生成された文書の部分集合の中から所望の部分集合を選択するクラスタ選択指示手段と、

前記クラスタ選択指示手段により選択された文書の部分集合を分類体系の構成要素として記憶する分類体系記憶手段と、

を備えたことを特徴とする文書分類装置。

【請求項９】　前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、

前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトルを、前記クラスタ選択指示手段により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正するベクトル修正手段と、

を備え、

前記分類手段は、前記ベクトル修正手段により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする請求項８に記載の文書分類装置。

【請求項１０】　前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、

前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正手段と、

を備え、

前記分類手段は、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項８に記載の文書分類装置。

【請求項１１】　前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、

前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正手段と、

を備え、

前記分類手段は、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル修正手段により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項９に記載の文書分類装置。

【請求項１２】　前記分類手段により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与手段を備え、

前記表示手段は、前記クラスタ特徴を表示するとともに、前記選択情報付与手段により付与された選択情報を表示することを特徴とする請求項８または１０に記載の文書分類装置。

【請求項１３】　前記分類体系記憶手段は、前記選択指示手段により選択された文書の部分集合に属する全部あるいは一部の文書のほか、クラスタ特徴および／または操作者が作成した任意の情報を分類体系の構成要素として記憶することを特徴とする請求項８～１２に記載の文書分類装置。

【請求項１４】　文書の内容にしたがって文書群を分類する文書分類装置において、

文書データ群を入力する文書入力手段と、

入力された文書データ群の各文書に対して所定の基準に基づき文書の分割をおこない、一つの文書データから一つまたは複数の分割文書データを生成する文書分割手段と、

前記文書データと前記分割文書データとの対応を示す文書－分割文書対応マップを生成する文書－分割文書対応マップ生成手段と、

前記分割文書データを分類する分割文書分類手段と、

前記分割文書分類手段による分類結果に基づいて分割文書分類結果情報を生成

する分割文書分類結果生成手段と、

前記文書－分割文書対応マップと前記分割文書分類結果情報とを用いて前記文書データの分類結果情報を生成する文書分類結果生成手段と、

を備えたことを特徴とする文書分類装置。

【請求項１５】　前記文書データを保存する文書保存手段と、

前記分割文書データを保存する分割文書保存手段と、

前記文書－分割文書対応マップ生成手段により生成された文書－分割文書対応マップを保存する文書－分割文書対応マップ保存手段と、

を備えたことを特徴とする請求項１４に記載の文書分類装置。

【請求項１６】　前記分割文書分類結果生成手段により生成された分割文書分類結果情報を保存する分割文書分類結果保存手段を備えたことを特徴とする請求項１５に記載の文書分類装置。

【請求項１７】　前記文書分割手段により生成される複数の分割文書データには分割前の文書データそのものを含むことを特徴とする請求項１４～１６のいずれか一つに記載の文書分類装置。

【請求項１８】　前記文書分割手段が、文書データの構造情報を基に文書データを分割する構成にしたことを特徴とする請求項１４～１７のいずれか一つに記載の文書分類装置。

【請求項１９】　前記文書データに含まれる要素を抽出する文書要素抽出手段と、

前記文書要素抽出手段により抽出された要素に付随する要素付随情報を抽出する要素付随情報抽出手段と、

を備え、

前記文書分割手段が、前記文書要素抽出手段により抽出された要素、または前記要素と前記要素付随情報抽出手段により抽出された要素付随情報とを用いて前記文書データを分割する構成にしたことを特徴とする請求項１４～１７のいずれか一つに記載の文書分類装置。

【請求項２０】　前記文書分割手段が、指示された指定範囲にしたがって文書データの分割をおこなう構成にしたことを特徴とする請求項１４～１７のいず

れか一つに記載の文書分類装置。

【請求項２１】　前記文書分割手段が、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する構成にしたことを特徴とする請求項１４～１７のいずれか一つに記載の文書分類装置。

【請求項２２】　前記文書分類結果生成手段が、文書データを示す情報および前記文書データに付随する代表的情報を、分類結果情報として抽出して提示する構成にしたことを特徴とする請求項１４～２１のいずれか一つに記載の文書分類装置。

【請求項２３】　前記文書分類結果生成手段が、分割文書データを示す情報および前記分割文書データに付随する代表的情報を、分類結果情報として、抽出して提示する構成にしたことを特徴とする請求項２２に記載の文書分類装置。

【請求項２４】　入力された複数の文書データを所定の形式で表示または印刷するために出力する文書処理方法において、

入力された文書データを記憶する文書記憶工程と、

前記文書記憶工程により記憶された文書データの全部または一部を選択する選択工程と、

前記選択工程により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出工程と、

前記特徴抽出工程により抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理工程と、

前記加工処理工程により加工処理された文書データの全部または一部を出力する出力工程と、

を含んだことを特徴とする文書処理方法。

【請求項２５】　前記出力工程は、

前記加工処理工程により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定工程と、

前記項目値設定工程により設定された項目値ごとに前記文書データの全部または一部を集計する集計工程と、

を含み、

前記文書データの全部または一部を、項目値を少なくとも一つの軸とする表形式に展開して出力することを特徴とする請求項２４に記載の文書処理方法。

【請求項２６】　前記出力工程は、さらに、前記加工処理工程により加工処理された文書データの全部または一部を、前記加工処理工程により加工処理される前の文書データの全部または一部とともに出力することを特徴とする請求項２４または２５に記載の文書処理方法。

【請求項２７】　前記文書記憶工程は、さらに、前記加工処理工程により加工処理された文書データの全部または一部を記憶することを特徴とする請求項２４～２６のいずれか一つに記載の文書処理方法。

【請求項２８】　前記選択工程は、さらに、前記出力工程により出力された文書データの全部または一部を選択することを特徴とする請求項２４～２７のいずれか一つに記載の文書処理方法。

【請求項２９】　前記文書記憶工程は、さらに、前記加工処理の内容に関するデータを記憶することを特徴とする請求項２４～２８のいずれか一つに記載の文書処理方法。

【請求項３０】　文書の内容に基づいて文書の分類をおこなう文書分類方法において、

文書データを入力する入力工程と、

前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、

前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、

前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、

前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、

前記クラスタ特徴算出工程により算出されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、

を含んだことを特徴とする文書分類方法。

【請求項３１】　文書の内容に基づいて文書の分類をおこなう文書分類方法において、

文書データを入力する入力工程と、

前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、

前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、

前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、

前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、

前記クラスタ特徴算出工程により算出されたクラスタ特徴を表示する表示工程と、

前記分類工程により生成された文書の部分集合の中から所望の部分集合を選択するクラスタ選択指示工程と、

前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、

を含んだことを特徴とする文書分類方法。

【請求項３２】　前記ベクトル生成工程により生成された文書特徴ベクトルを、前記クラスタ選択指示工程により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正するベクトル修正工程と、

を含み、

前記分類工程は、前記ベクトル修正工程により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする請求項３１に記載の文書分類方法。

【請求項３３】　前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正工程と

、

を含み、

前記分類工程は、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル生成手段工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項３１に記載の文書分類方法。

【請求項３４】　前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正工程と、

を含み、

前記分類工程は、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル修正工程により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする請求項３２に記載の文書分類方法。

【請求項３５】　前記分類工程により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与工程を含み、

前記表示工程は、前記クラスタ特徴を表示するとともに、前記選択情報付与工程により付与された選択情報を表示することを特徴とする請求項３１または３３に記載の文書分類方法。

【請求項３６】　前記分類体系生成工程は、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報に基づいて分類体系の構成要素を生成することを特徴とする請求項３１～３５に記載の文書分類方法。

【請求項３７】　文書の内容にしたがって文書群を分類する文書分類方法において、

文書データ群を入力し、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割をおこない、一つの文書データから一つまたは複数の分割文書データを生成し、前記文書データと前記分割文書データとの対応を示す文書－分割文書対応マップを生成し、前記分割文書データを分類し、分割文書分類結果

情報を生成し、前記文書－分割文書対応マップと前記分割文書分類結果情報とを用いて前記文書データの分類結果情報を生成することを特徴とする文書分類方法。

【請求項３８】　前記請求項２４～３７のいずれか一つに記載された方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

この発明は、入力された複数の文書データを所定の形式で表示または印刷するために出力する文書処理装置、文書処理方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。また、この発明は、入力された複数の文書をその文書の内容に基づいて分類をおこなう、特に文書分類の際に算出される分類カテゴリ（体系）を精錬化する文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。

【０００２】

【従来の技術】

近年、さまざまな文書分類装置や文書検索装置が開発されている。また、インターネット等のネットワーク技術の普及により国内外の大量の電子化文書へのアクセスが可能になり、それに比例して業務上電子的に蓄積される情報の量も飛躍的に拡大した。その中で収集した大量の文書情報を意味あるカテゴリ（体系）に分類する等の知的作業の必要性が高まってきている。

【０００３】

これらの大量の文書情報を意味的に分類するという作業の目的は、以下のようなものである。まず第１に、検索容易性の向上が考えられる。これは、膨大な文書群を分類名称（内容名）を手がかりに検索できるので検索が比較的容易になるというものである。

【０００４】

第２に、情報群全体の把握が考えられる。これは、文書群全体がどのような内容（個々の分類）で構成されているかを把握する。しかし、大量の文書情報を操作者が手動で分類する場合、正確な分類をすることはできるが、分類に係る人的・時間的コストが膨大なものになるため、近年の文書の蓄積量の膨大さから、文書情報の自動分類装置が提案されるようになってきた。

【０００５】

文書自動分類装置の従来技術としては、たとえば、特開平７－３６８９７号公報に記載されているように、文書を、単語を特徴とする文書ベクトルとみなし、クラスタリング手法を用いてこれらの文書ベクトルを群分けし、群分けした文書ベクトルに基づいて文書の自動分類をおこなうものがある。

【０００６】

また、「Ｐｒｏｊｅｃｔｉｏｎｓ ｆｏｒ Ｅｆｆｉｃｉｅｎｔ Ｄｏｃｕｍｅｎｔ Ｃｌｕｓｔｅｒｉｎｇ（著者名：Ｈｉｎｒｉｃｈ Ｓｃｈｕｔｚｅ ａｎｄ Ｃｒａｉｎｇ Ｓｉｌｖｅｒｓｔｅｉｎ，　学会名：ＡＣＭ，　論文名：Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ＳＩＧＩＲ，　ページ：７４－８１，　発行年：１９９７）」においては、潜在的意味空間において文書分類を実施しているものがある。そのほかの方法としては、確率論的アプローチを用いる方法等が考えられる。

【０００７】

また近年、インターネットなどの普及により、大量の文書群へのアクセスが可能になり、その結果、その文書群をさまざまな利用者の意図に基づいて、かつ、効率的に利用できるようにする必要性が高まっている。そのため、大量の文書群を意味のあるカテゴリに分類し、文書群の構造を把握するという知的作業がおこなわれ始めている。しかし、このような分類作業を人手によりおこなう場合、その人的および時間的なコストが膨大なものになるし、また、分類のための知識を分類者のみが有することになるため、分類担当者が代わると分類基準も変わってしまうことになる。

【０００８】

そのため、文書群を人間が分類するような分類基準で自動的に分類しうる文書

分類装置が望まれており、文書分類装置としては、たとえば、特開平７－１１４５７２号公報に記載されているように、文書から自動的に単語の特徴ベクトルを抽出し、その特徴ベクトルをもとに文書分類することで、意味的な異なりを用いた自動分類を可能にするものがある。

【０００９】

【発明が解決しようとする課題】

しかしながら、上記従来技術の文書分類装置は、本質的には単語で構成される多次元空間に布置した文書を統計的な分類をする方法であるため、分類結果は単語のいわゆる振る舞いという観点から統計的に求められたものにすぎず，分類の結果、算出される各クラスタ（分類された個々の文書の部分集合）が操作者（利用者）に理解不能な場合がある。

【００１０】

また、どのような分類結果が最適かは、分類対象の文書集合の特徴や、利用者の作業の目的に依存するため、最適な分類結果について定義することが困難であるという問題点があった。特に、上記情報群全体の把握に関し、多様な操作者の意図により要求される分類も異なるため、一度の分類作業で、操作者の所望する結果を得ることが困難であるという問題点があった。

【００１１】

このように、文書分類の結果は、多くのいわゆるノイズを含んだものであると解釈することができ、その一部についてのみが操作者にとって有益な場合が多いという問題点があった。

【００１２】

また、これらの従来技術においては、文書の構成単位を考慮していないため、文書が一つまたは複数の段落記号やタイトルなどにより区切られた構造を持つ場合には、一つの文書の中に複数の話題や意味が含まれてしまい、その結果、利用者がその分類カテゴリを理解し難くなったり、また、ある特定の話題や特定の意味に限定されたカテゴリになったり、利用者の意図するカテゴリとは異なるカテゴリに分類されてしまうという問題か生じている。

【００１３】

なお、特開平６－１７６０６４号公報に示された文脈依存自動分類装置には、文書の段落情報を考慮した文書自動分類をおこなうことにより分類精度を高めようとするものが開示されているか、本質的に上記の問題を解決するものではない。

【００１４】

また、上記従来技術の文書分類装置や文書検索装置等の文書処理装置は、単に文書を分類する、あるいは文書を検索する機能を有するのみで、その結果を用いてさらなる分析をおこない、文書群に内在する隠れた情報の解析をおこなうことについては何ら考慮がされておらず、文書群に内在する隠れた情報の解析は別の解析装置を用いておこなわなければならないという問題点があった。

【００１５】

また、情報分析をおこなう操作者が分類作業や検索作業をおこなうのは、これらの作業において、結果は目的なのではなく、単に情報分析作業の途中経過にすぎないからである。通常は、その後、さらに結果を把握しやすくするために、元の文書に含まれる情報を最大限に活用し、結果の並べ替えをおこなったり、集計・統計処理を施したり、結果をもとに表の形式にまとめたり、さらにはグラフ化したりというようなさまざまな処理を繰り返しおこない、意味ある情報分析結果を導き出す必要がある。

【００１６】

また、数値データを対象とする情報の分析作業において、表計算ソフトウエアが用いられる場合があるが、表計算ソフトウエアは、元来、数値データの取扱いを意図して開発されたものであり、文字データ、特に文書の意味に係わるような分析作業においては十分な効果を発揮することはできなかった。

【００１７】

この発明は、上述した従来例による問題点を解消するため、文書の意味に係わるような分析作業において、単に分類作業や検索作業などを固定された機能としておこない、その結果を出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる文書処理装置、文書処理方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体を提

供することを第１の目的とする。

【００１８】

またこの発明は、上述した従来例による問題点を解消するため、任意の文書集合にどのような内容が含まれるかを漸次的に収集することができる文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体を提供することを第２の目的とする。

【００１９】

またこの発明は、上述した従来例による問題点を解決するため、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されることがないことにより、利用者かその分類カテゴリをよく理解できる文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体を提供することを第３の目的とする。

【００２０】

【課題を解決するための手段】

上述した課題を解決し、目的を達成するため、請求項１の発明に係る文書処理装置は、入力された複数の文書データを所定の形式で表示または印刷するために出力する文書処理装置において、入力された文書データを記憶する文書記憶手段と、前記文書記憶手段により記憶された文書データの全部または一部を選択する選択手段と、前記選択手段により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出手段と、前記特徴抽出手段により抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理手段と、前記加工処理手段により加工処理された文書データの全部または一部を出力する出力手段と、を備えたことを特徴とする。

【００２１】

この請求項１の発明によれば、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【００２２】

また、請求項２の発明に係る文書処理装置は、請求項１の発明において、前記出力手段が、前記加工処理手段により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定手段と、前記項目値設定手段により設定された項目値ごとに前記文書データの全部または一部を集計する集計手段と、を備え、前記文書データの全部または一部を、項目値を少なくとも一つの軸とする表形式に展開して出力することを特徴とする。

【００２３】

この請求項２の発明によれば、簡易な操作で加工処理の結果をクロス表として表すことができ、情報の内容の把握を容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【００２４】

また、請求項３の発明に係る文書処理装置は、請求項１または２の発明において、前記出力手段が、さらに、前記加工処理手段により加工処理された文書データの全部または一部を、前記加工処理手段により加工処理される前の文書データの全部または一部とともに出力することを特徴とする。

【００２５】

この請求項３の発明によれば、加工処理すべき対象データとその他のデータが同時に表示され、それを確認することにより、加工処理の対象範囲の決定を正確かつ容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【００２６】

また、請求項４の発明に係る文書処理装置は、請求項１～３の発明において、前記文書記憶手段が、さらに、前記加工処理手段により加工処理された文書データの全部または一部を記憶することを特徴とする。

【００２７】

この請求項４の発明によれば、以後、他のデータと同様に扱うことができるこ

とから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【００２８】

また、請求項５の発明に係る文書処理装置は、請求項１～４の発明において、前記選択手段が、さらに、前記出力手段により出力された文書データの全部または一部を選択することを特徴とする。

【００２９】

この請求項５の発明によれば、出力手段により出力された文書データの全部または一部をさらなる分析の対象とすることができ、多彩で高度な情報分析作業ができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【００３０】

また、請求項６の発明に係る文書処理装置は、請求項１～５の発明において、前記文書記憶手段が、さらに、前記加工処理の内容に関するデータを記憶することを特徴とする。

【００３１】

この請求項６の発明によれば、加工処理の内容に関するデータの紛失を防止し、当該データの管理が容易になるだけでなく、加工処理に用いた設定とそれによる処理結果を関連づけて把握することができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【００３２】

また、請求項７の発明に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、文書データを入力する入力手段と、前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の

部分集合を生成する分類手段と、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、前記クラスタ特徴算出手段により算出されたクラスタ特徴を分類体系の構成要素として記憶する分類体系記憶手段と、を備えたことを特徴とする。

【００３３】

この請求項７の発明によれば、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができる。

【００３４】

また、請求項８の発明に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、文書データを入力する入力手段と、前記入力手段により入力された文書データを解析して言語解析情報を得る言語解析手段と、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成手段と、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類手段と、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出手段と、前記クラスタ特徴算出手段により算出されたクラスタ特徴を表示する表示手段と、前記分類手段により生成された文書の部分集合の中から所望の部分集合を選択するクラスタ選択指示手段と、前記クラスタ選択指示手段により選択された文書の部分集合を分類体系の構成要素として記憶する分類体系記憶手段と、を備えたことを特徴とする。

【００３５】

この請求項８の発明によれば、選択されたクラスタのみを用いており、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【００３６】

また、請求項９の発明に係る文書分類装置は、請求項８の発明において、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクト

ルを、前記クラスタ選択指示手段により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正するベクトル修正手段と、を備え、前記分類手段が、前記ベクトル修正手段により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする。

【００３７】

この請求項９の発明によれば、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【００３８】

また、請求項１０の発明に係る文書分類装置は、請求項８の発明において、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正手段と、を備え、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【００３９】

この請求項１０の発明によれば、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【００４０】

また、請求項１１の発明に係る文書分類装置は、請求項９の発明において、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶する文書特徴ベクトル記憶手段と、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正手段と、を備え、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル修正手段により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【００４１】

この請求項１１の発明によれば、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【００４２】

また、請求項１２の発明に係る文書分類装置は、請求項８または１０の発明において、前記分類手段により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与手段を備え、前記表示手段が、前記クラスタ特徴を表示するとともに、前記選択情報付与手段により付与された選択情報を表示することを特徴とする。

【００４３】

この請求項１２の発明によれば、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができる。

【００４４】

また、請求項１３の発明に係る文書分類装置は、請求項８～１２の発明において、前記分類体系記憶手段が、前記選択指示手段により選択された文書の部分集合に属する全部あるいは一部の文書のほか、クラスタ特徴および／または操作者が作成した任意の情報を分類体系の構成要素として記憶することを特徴とする。

【００４５】

この請求項１３の発明によれば、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できるので、分類体系の利用価値を向上させることができる。

【００４６】

また、請求項１４の発明に係る文書分類装置は、文書の内容にしたがって文書群を分類する文書分類装置において、文書データ群を入力する文書入力手段と、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割をおこない、一つの文書データから一つまたは複数の分割文書データを生成する文書分割手段と、前記文書データと前記分割文書データとの対応を示す文書－分割文書

対応マップを生成する文書－分割文書対応マップ生成手段と、前記分割文書データを分類する分割文書分類手段と、前記分割文書分類手段による分類結果に基づいて分割文書分類結果情報を生成する分割文書分類結果生成手段と、前記文書－分割文書対応マップと前記分割文書分類結果情報とを用いて前記文書データの分類結果情報を生成する文書分類結果生成手段と、を備えたことを特徴とする。

【００４７】

この請求項１４の発明によれば、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりすることがなく、したがって、利用者がその分類カテゴリをよく理解できる。また、分割前文書（所属文書）中の分割文書の位置なども示されるので、利用者は文書群中の読みたい部分を効率的に読むことができる。

【００４８】

また、請求項１５の発明に係る文書分類装置は、請求項１４の発明において、前記文書データを保存する文書保存手段と、前記分割文書データを保存する分割文書保存手段と、前記文書－分割文書対応マップ生成手段により生成された文書－分割文書対応マップを保存する文書－分割文書対応マップ保存手段と、を備えたことを特徴とする。

【００４９】

この請求項１５の発明によれば、分割文書データおよび文書－分割文書対応マップを再生成することなしに、同一の文書データに対して、分類数、分類手法、または分類時の諸設定などパラメータの異なる分類結果を効率的に求めることができる。また、文書データを分類し、分類結果を生成するために必要なデータか保存されることにより、利用者が分類作業に対して時間的な自由度を持つことができるし、過去に行った文書分類の再分析を任意の時間間におこなうこともできる。

【００５０】

また、請求項１６の発明に係る文書分類装置は、請求項１５の発明において、前記分割文書分類結果生成手段により生成された分割文書分類結果情報を保存す

る分割文書分類結果保存手段を備えたことを特徴とする。

【００５１】

この請求項１６の発明によれば、請求項１５の発明の効果に加え、一度分類を実行すれば、その分類結果をテキスト表現や表表現やグラフ表現などさまざまな形式で表現することができる。また、分割文書分類結果情報が保存されることにより、分類の実行作業および分類結果の分析作業において、利用者が時間的な自由度を持つことができるし、過去に行った文書分類結果の再分析をさまざまな表現形式で任意の時間におこなうこともできる。

【００５２】

また、請求項１７の発明に係る文書分類装置は、請求項１４～１６の発明において、前記文書分割手段により生成される複数の分割文書データには分割前の文書データそのものを含むことを特徴とする。

【００５３】

この請求項１７の発明によれば、利用者は、分割されている文書データを分類することで得られる詳細な文書データの分類構造だけでなく、分割前の文書データ自体を分類した結果として得られる概略的でマクロな分類構造の融合した分類構造を得ることができる。

【００５４】

また、請求項１８の発明に係る文書分類装置は、請求項１４～１７の発明において、前記文書分割手段が、文書データの構造情報を基に文書データを分割する構成にしたことを特徴とする。

【００５５】

この請求項１８の発明によれば、異なった話題の分割等を適切におこなうことができ、したがって、文書データの詳細な分類構造がわかる文書分類を適切におこなうことができる。

【００５６】

また、請求項１９の発明に係る文書分類装置は、請求項１４～１７の発明において、前記文書データに含まれる要素を抽出する文書要素抽出手段と、前記文書要素抽出手段により抽出された要素に付随する要素付随情報を抽出する要素付随

情報抽出手段と、を備え、前記文書分割手段が、前記文書要素抽出手段により抽出された要素、または前記要素と前記要素付随情報抽出手段により抽出された要素付随情報とを用いて前記文書データを分割する構成にしたことを特徴とする。

【００５７】

この請求項１９の発明によれば、文書データの詳細な分類構造がわかる文書分類を適切におこなうことができる。

【００５８】

また、請求項２０の発明に係る文書分類装置は、請求項１４～１７の発明において、前記文書分割手段が、指示された指定範囲にしたがって文書データの分割をおこなう構成にしたことを特徴とする。

【００５９】

この請求項２０の発明によれば、利用者の意図に合い、かつ文書データの詳細な分類構造がわかる文書分類をおこなうことができる。

【００６０】

また、請求項２１の発明に係る文書分類装置は、請求項１４～１７において、前記文書分割手段が、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する構成にしたことを特徴とする。

【００６１】

この請求項２１の発明によれば、話題の異なった内容などが異なった文書として分類される可能性が高くなり、したがって、この発明でも文書データの詳細な分類構造がわかる文書分類をおこなうことができる。

【００６２】

また、請求項２２の発明に係る文書分類装置は、請求項１４～２１の発明において、前記文書分類結果生成手段が、文書データを示す情報および前記文書データに付随する代表的情報を、分類結果情報として抽出して提示する構成にしたことを特徴とする。

【００６３】

この請求項２２の発明によれば、利用者は文書データの詳細な分類構造の概要や全体的な構造を容易に把握することができる。

【００６４】

また、請求項２３の発明に係る文書分類装置は、請求項２２の発明において、前記文書分類結果生成手段が、分割文書データを示す情報および前記分割文書データに付随する代表的情報を、分類結果情報として、抽出して提示する構成にしたことを特徴とする。

【００６５】

この請求項２３の発明によれば、利用者は文書データの詳細な分類構造の概要や全体的な構造とともにどの分割文書が起因して当該カテゴリに分類されたかというようなことも容易にわかる。

【００６６】

また、請求項２４の発明に係る文書処理方法は、入力された複数の文書データを所定の形式で表示または印刷するために出力する文書処理方法において、入力された文書データを記憶する文書記憶工程と、前記文書記憶工程により記憶された文書データの全部または一部を選択する選択工程と、前記選択工程により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出工程と、前記特徴抽出工程により抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理工程と、前記加工処理工程により加工処理された文書データの全部または一部を出力する出力工程と、を含んだことを特徴とする。

【００６７】

この請求項２４の発明によれば、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【００６８】

また、請求項２５の発明に係る文書処理方法は、請求項２４の発明において、前記出力工程が、前記加工処理工程により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定工程と、前記項目値設定工程により設定された項目値ごとに前記文書データの全部または一部を集計する集計工程と、を含み、前記文書データの全部または一部を、項目値を少なく

とも一つの軸とする表形式に展開して出力することを特徴とする。

【００６９】

この請求項２５の発明によれば、簡易な操作で加工処理の結果をクロス表として表すことができ、情報の内容の把握を容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【００７０】

また、請求項２６の発明に係る文書処理方法は、請求項２４または２５の発明において、前記出力工程が、さらに、前記加工処理工程により加工処理された文書データの全部または一部を、前記加工処理工程により加工処理される前の文書データの全部または一部とともに出力することを特徴とする。

【００７１】

この請求項２６の発明によれば、加工処理すべき対象データとその他のデータが同時に表示され、それを確認することにより、加工処理の対象範囲の決定を正確かつ容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【００７２】

また、請求項２７発明に係る文書処理方法は、請求項２４～２６の発明において、前記文書記憶工程が、さらに、前記加工処理工程により加工処理された文書データの全部または一部を記憶することを特徴とする。

【００７３】

この請求項２７の発明によれば、以後、他のデータと同様に扱うことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【００７４】

また、請求項２８の発明に係る文書処理方法は、請求項２４～２７の発明において、前記選択工程が、さらに、前記出力工程により出力された文書データの全部または一部を選択することを特徴とする。

【００７５】

この請求項２８の発明によれば、出力手段により出力された文書データの全部または一部をさらなる分析の対象とすることができ、多彩で高度な情報分析作業ができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【００７６】

また、請求項２９の発明に係る文書処理方法は、請求項２４～２８の発明において、前記文書記憶工程が、さらに、前記加工処理の内容に関するデータを記憶することを特徴とする。

【００７７】

この請求項２９の発明によれば、加工処理の内容に関するデータの紛失を防止し、当該データの管理が容易になるだけでなく、加工処理に用いた設定とそれによる処理結果を関連づけて把握することができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことができる。

【００７８】

また、請求項３０の発明に係る文書分類方法は、文書の内容に基づいて文書の分類をおこなう文書分類方法において、文書データを入力する入力工程と、前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、前記クラスタ特徴算出工程により算出されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、を含んだことを特徴とする。

【００７９】

この請求項３０の発明によれば、クラスタを得ることができるとともに、クラ

スタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができる。

【００８０】

また、請求項３１の発明に係る文書分類方法は、文書の内容に基づいて文書の分類をおこなう文書分類方法において、文書データを入力する入力工程と、前記入力工程により入力された文書データを解析して言語解析情報を得る言語解析工程と、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するベクトル生成工程と、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成する分類工程と、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出するクラスタ特徴算出工程と、前記クラスタ特徴算出工程により算出されたクラスタ特徴を表示する表示工程と、前記分類工程により生成された文書の部分集合の中から所望の部分集合を選択するクラスタ選択指示工程と、前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成する分類体系生成工程と、を含んだことを特徴とする。

【００８１】

この請求項３１の発明によれば、選択されたクラスタのみを用いており、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【００８２】

また、請求項３２の発明に係る文書分類方法は、請求項３１の発明において、前記ベクトル生成工程により生成された文書特徴ベクトルを、前記クラスタ選択指示工程により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正するベクトル修正工程と、を含み、前記分類工程が、前記ベクトル修正工程により修正された文書特徴ベクトルに基づいて文書を分類することを特徴とする。

【００８３】

この請求項３２の発明によれば、既知になったクラスタの影響を排除した新た

なクラスタを生成することができる。

【００８４】

また、請求項３３の発明に係る文書分類方法は、請求項３１の発明において、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正工程と、を含み、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル生成手段工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【００８５】

この請求項３３の発明によれば、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【００８６】

また、請求項３４の発明に係る文書分類方法は、請求項３２の発明において、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特徴量に基づいて修正する文書表現空間修正工程と、を含み、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル修正工程により修正された文書特徴ベクトル間の類似度に基づいて文書を分類することを特徴とする。

【００８７】

この請求項３４の発明によれば、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【００８８】

また、請求項３５の発明に係る文書分類方法は、請求項３１または３３の発明において、前記分類工程により生成された文書の部分集合に所属する文書のすべ

てあるいは一部が選択された場合に選択されたことを示す選択情報を付与する選択情報付与工程を含み、前記表示工程が、前記クラスタ特徴を表示するとともに、前記選択情報付与工程により付与された選択情報を表示することを特徴とする。

【００８９】

この請求項３５の発明によれば、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができる。

【００９０】

また、請求項３６の発明に係る文書分類方法は、請求項３１～３５の発明において、前記分類体系生成工程が、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報に基づいて分類体系の構成要素を生成することを特徴とする。

【００９１】

この請求項３６の発明によれば、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できるので、分類体系の利用価値を向上させることができる。

【００９２】

また、請求項３７の発明に係る文書分類方法は、文書の内容にしたがって文書群を分類する文書分類方法において、文書データ群を入力し、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割をおこない、一つの文書データから一つまたは複数の分割文書データを生成し、前記文書データと前記分割文書データとの対応を示す文書－分割文書対応マップを生成し、前記分割文書データを分類し、分割文書分類結果情報を生成し、前記文書－分割文書対応マップと前記分割文書分類結果情報とを用いて前記文書データの分類結果情報を生成することを特徴とする。

【００９３】

この請求項３７の発明によれば、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利

用者の意図するカテゴリとは異なるカテゴリに分類されたりすることがなく、したがって、利用者がその分類カテゴリをよく理解できる。また、分割前文書（所属文書）中の分割文書の位置なども示されるので、利用者は文書群中の読みたい部分を効率的に読むことができる。

【００９４】

また、請求項３８の発明に係る記憶媒体は、請求項２４～３７に記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項２４～３７の動作をコンピュータによって実現することが可能である。

【００９５】

【発明の実施の形態】

以下に添付図面を参照して、この発明に係る文書処理装置、文書処理方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体の好適な実施の形態を詳細に説明する。

【００９６】

〔実施の形態１〕

まず、この発明の実施の形態１による文書処理装置を構成する情報処理システム全体のハードウエア構成を説明する。図１は、実施の形態１による文書処理装置を構成する情報処理システム全体のハードウエア構成を示す説明図である。

【００９７】

図１において、実施の形態１による文書処理装置を構成する情報処理システムは、サーバー／クライアント方式で構成されている。すなわち、サーバー１０１と複数のクライアント１０２がネットワーク１０３によって接続されている。クライアント１０２は、分類データ等の加工データの生成、サーバー１０１への指示、分類結果等の加工処理結果の表示などをおこなう。一方、クライアント１０２からの指示にしたがって、サーバー１０１は文書（テキスト）分類等の加工処理を膨大な数値演算によりおこない、その処理の結果をクライアント１０２へ送る。

【００９８】

分類処理の場合、より具体的には、サーバー１０１においては、テキスト分類処理（前処理、クラスタリング処理）がおこなわれ、クライアント１０２においては、分類データ生成、処理実行指示、テキスト分類結果表示等がおこなわれる。サーバー１０１における処理は、上述のように、「前処理」と「分類処理」の二つに分かれており、その処理はデータによっては非常に負荷が大きくなる。したがって、サーバー１０１は「前処理」と「分類処理」がそれぞれ一つずつしか処理をおこなわないようにマネージャプロセスが処理受付リストを作成して管理する。

【００９９】

また、サーバー１０１とクライアント１０２との間のデータのやりとりはファイル共有という方法を用いる。すなわち、分類処理等の加工処理に用いるファイルをサーバー１０１上の共有フォルダに作成することにより両者はデータのやりとりをおこなう。したがって、クライアント１０２からはサーバー１０１の共有フォルダをネットワーク共有して利用することが可能である。

【０１００】

つぎに、サーバー１０１およびクライアント１０２のハードウエア構成について説明する。図２は、実施の形態１による文書処理装置を構成する情報処理システムにおけるサーバー１０１のハードウエア構成を示す説明図である。サーバー１０１は、たとえばワークステーション（ＷＳ）等が用いられる。

【０１０１】

図２において、２０１はサーバー１０１全体を制御するＣＰＵを、２０２はブートプログラム等を記憶したＲＯＭを、２０３はＣＰＵ２０１のワークエリアとして使用されるＲＡＭ２０３を、２０４は通信回線２０５を介してネットワーク１０３に接続され、そのネットワーク１０３と内部のインターフェイスを司るインターフェイス（Ｉ／Ｆ）を、２０６はデータを記憶するディスク装置を示している。２００は上記各部を結合させるためのバスを示している。

【０１０２】

そのほか、文書情報、画像情報、機能情報等を表示するディスプレイ２０８や、データを入力するためのキーボード２０９およびマウス２１０等が同様に接続

されていてもよい。さらに、ディスク装置２０６には、クライアント１０２との間のデータのやりとりをするための共有フォルダ２０７が設けられている。

【０１０３】

また、図３は、実施の形態１による文書処理装置を構成する情報処理システムにおけるクライアント１０２のハードウエア構成を示す説明図である。クライアント１０２は、たとえばパーソナルコンピュータ（ＰＣ）等が用いられる。

【０１０４】

図３において、３０１はシステム全体を制御するＣＰＵを、３０２はブートプログラム等を記憶したＲＯＭを、３０３はＣＰＵ３０１のワークエリアとして使用されるＲＡＭを、３０４はＣＰＵ３０１の制御にしたがってＨＤ（ハードディスク）３０５に対するデータのリード／ライトを制御するＨＤＤ（ハードディスクドライブ）を、３０５はＨＤＤ３０４の制御で書き込まれたデータを記憶するＨＤを、３０６はＣＰＵ３０１の制御にしたがってＦＤ（フロッピーディスク）３０７に対するデータのリード／ライトを制御するＦＤＤ（フロッピーディスクドライブ）を、３０７はＦＤＤ３０６の制御で書き込まれたデータを記憶する着脱自在のＦＤを、３０８はドキュメント、画像、機能情報等を表示するディスプレイをそれぞれ示している。

【０１０５】

また、３０９は通信回線３１０を介してネットワーク１０３に接続され、そのネットワーク１０３と内部のインターフェイスを司るインターフェイス（Ｉ／Ｆ）を、３１１は文字、数値、各種指示等の入力のためのキーを備えたキーボードを、３１２はカーソルの移動や範囲選択、あるいは表示画面に表示されたアイコンやボタンの押下やウインドウの移動やサイズの変更等をおこなうマウスを、３１３はＯＣＲ（Ｏｐｔｉｃａｌ　Ｃｈａｒａｃｔｅｒ　Ｒｅａｄｅｒ）機能を備えた画像を光学的に読み取るスキャナを、３１４は分類結果を含むデータの内容等を印刷するプリンタを、３１５は上記各部を結合するためのバスをそれぞれ示している。また、ＨＤ３０５にはワープロソフト等のアプリケーションソフト３１６が記憶されている。

【０１０６】

つぎに、実施の形態１による文書処理装置の機能的構成について説明する。図４は、実施の形態１による文書処理装置の構成を機能的に示すブロック図である。図４において、文書処理装置は、入力部４０１と、文書記憶部４０２と、選択部４０３と、特徴抽出部４０４と、加工処理部４０５と、出力部４０６を含む構成である。

【０１０７】

入力部４０１、文書記憶部４０２、選択部４０３、特徴抽出部４０４、加工処理部４０５、出力部４０６は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０１０８】

入力部４０１は、文書データを入力するものであり、たとえば、キーボード２０９または３１１、スキャナ３１３、ＯＣＲ機能を備えたスキャナ３１３、またはネットワーク１０３を経由して文書や文書群を得ることができるＩ／Ｆ２０４または３０９等である。また、入力部４０１は、上記以外に、文書データを取得することができるものであれば、それらのすべてを含む。たとえば、文書データがデータベース化されている場合に、そのデータベースが記録された媒体を実施の形態１の文書処理装置に組み入れた場合も文書データの入力とする。

【０１０９】

ここで、文書とは、自然言語で記述された一つ以上の文の集まりであり、文字、文字列、数値等から構成されており、それらの意味のあるまとまりを一つの文書とする。また、複数の文書の集まりを、文書群とする。

【０１１０】

文書は一つあるいは複数の項目から構成されている。項目は、項目名と、項目値から構成されている。項目名は項目の内容を示すラベルであり、文書に含まれていても含まれていなくてもよい。項目値は項目の実際の内容である。図５は、実施の形態１による文書処理装置の項目名と項目値の関係を示す説明図である。たとえば、一つの特許公報は一つの文書であり、特許公報を項目名と項目値によ

って表現すると、図５のようになる。

【０１１１】

入力部４０１によって取得された文書あるいは文書群は、それぞれの文書に一意な文書ＩＤが付与され、文書記憶部４０２により記憶される。図６は、実施の形態１による文書処理装置の文書記憶部４０２に記憶された文書のデータ構造を示す説明図である。各項目名あるいは項目値は、文書記憶部４０２のセル、すなわち一つの記憶単位に収納される。

【０１１２】

図６においては、一つのセルは３つの記憶領域から構成されており、第１番目の記憶領域６０１にはつぎのセルの文書記憶部４０２上の位置（番地）が記憶されている。第２番目の記憶領域６０２には、セルの属性値が記憶されている。

【０１１３】

セルの属性値としては、たとえば、「０」が「空」、「１」が「数値」、「２」が文字列・・・というように設定することができる。第３番目の記憶領域６０３には、セルの実際の内容、すなわち、項目名あるいは項目値等が格納される領域の先頭位置が記憶されている。

【０１１４】

セルの順序の並び替えや、セルの追加・削除は、第１番目の記憶領域６０１に記憶されたつぎのセルの位置を変更することにより、容易に実現することができる。また、セルの実際の内容は、セルの構造とは異なる別の領域に記憶されているので、たとえば、項目を変更した結果、あらかじめ確保された領域では収まり切れなくなった場合には、セルの構造自体には影響なく、別途大きな領域を確保してそこに項目値を記憶し、第３番目に記憶された記憶領域６０３の先頭位置を変更するだけでよい。

【０１１５】

図７は、実施の形態１による文書処理装置の文書記憶部４０２に記憶された文書の別のデータ構造を示す説明図である。図７において、一つのセルは二つの記憶領域を使用している。第１番目の記憶領域７０１には、セル属性値が記憶されている。第２番目の記憶領域７０２には、セルの実際の内容、すなわち項目名あ

るいは項目値などが格納される領域の先頭位置が記憶されている。

【０１１６】

つぎのセルは、文書記憶部４０２上でとなり合うつぎの記憶領域に記憶されている。このデータ構造では、セルの順序の並び替え、セルの追加・削除が発生した場合には、記憶内容の移動操作が必要となる。

【０１１７】

文書記憶部４０２は、通常高速に情報を扱える半導体メモリで構成されるが、磁気ディスクあるいは光ディスク等で構成される補助記憶装置を含んでいてもよい。

【０１１８】

文書記憶部４０２に記憶された文書あるいは文書群は、出力部４０６により表示される。実施の形態１においては、出力部４０６は、ＣＲＴディスプレイ、液晶ディスプレイ等から構成される。出力部４０６は、文書記憶部４０２に記憶された文書あるいは文書群の内容をセルと順次たどって読み出し、表の形式で表示または印刷する。

【０１１９】

また、出力部４０６は、表の形式で表示または印刷されたデータに基づいてグラフを描画するグラフ描画部４０７を含んでいてもよい。グラフ描画部４０７は、文書記憶部４０２に記憶された文書あるいは文書群の項目値に対して利用者が設定した領域の内容を読み出し、利用者の指示により棒グラフ、円グラフ、折れ線グラフ等のグラフを描画し、表示または印刷する。

【０１２０】

出力部４０６は、入力部４０１による操作に関する表示、たとえば、操作メニューやマウスポインタ、カーソルの表示等もおこなう。また、処理結果を印刷するためのプリンタ等の印刷装置を含んでいてもよい。

【０１２１】

選択部４０３は、入力部４０１による操作者の指示により、出力部４０６の表示上で選択された領域のデータを文書記憶部４０２から読み出し、特徴抽出部４０４へ送る。選択部４０３の選択方法について、図８～図１０を用いて説明する

【０１２２】

図８～図１０は、実施の形態１による文書処理装置の出力部４０６による画面表示の例、具体的には、自動車の故障状況の内容が表示された画面表示の例を示す説明図である。図８において、画面表示には、文書ＩＤ番号を示す「番号」欄８０１、故障情報を受け付けた日付を示す「受付日」欄８０２、故障情報を受け付けた営業所を示す「営業所」欄８０３、故障情報の対象となった自動車の車種を示す「車種」欄８０４、故障情報対象となった自動車の年式を示す「年式」欄８０５、故障状況の内容を示す「内容」欄８０６が表示される。

【０１２３】

図９において、選択領域９０１は、矩形で囲まれ、表示色が変更されている部分であり、図１０においても同様に、選択領域１００１は、矩形で囲まれ、表示色が変更されている部分である。

【０１２４】

選択部４０３が選択する領域としては、図９に示すように、画面上の列の一部であってもよいし、また、図１０に示すように項目名を選択した場合はその項目名に属する項目値全部が選択されるようにしてもよい。なお、実施の形態１では、文字列の属性を持つ領域のみ選択可能とする。

【０１２５】

つぎに、特徴抽出部４０４によりおこなわれる抽出処理の内容について説明する。選択部４０３により選択された項目値は、特徴抽出部４０４によりその項目値の特徴が抽出される。図１１は、実施の形態１による文書処理装置の特徴抽出部４０４によりおこなわれる抽出処理の内容の一覧を示す説明図である。

【０１２６】

図１１において、抽出処理には、対象とする文字列に含まれる単語、その単語の単語数、単語の文字数、単語のそれぞれの出現回数．．．等がある。これらの抽出処理は、規則音声合成装置や自動翻訳装置等の一般的に用いられている形態素解析技術あるいは構文解析技術等の自然言語処理技術を用いて実現する。

【０１２７】

つぎに、加工処理部４０５によりおこなわれる加工処理の内容について説明する。特徴抽出部４０４により抽出処理された特徴量に対して、加工処理部４０５により加工処理が施される。図１２は、実施の形態１による文書処理装置の加工処理部４０５によりおこなわれる加工処理の内容の一覧を示す説明図である。

【０１２８】

加工処理には、同一の特徴量ごと分類する「分類処理」、所定の特徴量を検索する「検索処理」、特徴量の内容ごとに並べ替えをおこなう「並べ替え処理」、特徴量の代表値を抽出する「代表値抽出処理」、特徴量のうちの最大値を抽出する「最大値抽出処理」、特徴量のうち最小値を抽出する「最小値抽出処理」、特徴量を算術する「算術処理」等がある。

【０１２９】

特徴抽出部４０４によりおこなわれる特徴量の抽出処理の内容と、加工処理部４０５によりおこなわれる抽出された特徴量の加工処理の内容の組み合わせは、おのおの操作者が選択できるようにすることができる。また、効果の高い組み合わせをあらかじめ設定して、その設定された組み合わせを操作者に提供するようにしてもよい。

【０１３０】

加工処理部４０５により加工処理された処理結果は、加工処理部４０５内の加工処理結果保持部４０８に保持される。加工処理結果保持部４０８に保持された加工処理結果は、出力部４０６により出力される。出力部４０６は、加工処理結果保持部４０８から内容を読み出し、画像表示や印刷出力をおこなう。

【０１３１】

ここで、特徴抽出部４０４により抽出される特徴（量）として、項目値に含まれる単語それぞれの出現回数を選択し、加工処理部４０５によりおこなわれる加工処理として、分類処理を選択した場合について説明する。

【０１３２】

一般的に、二つの文書があり、それら二つの文書を構成する単語の出現頻度が等しい場合、それら二つの文書の意味は似通っていると考えることができる。すなわち、ある文書での単語の出現回数は、その文書の意味に関係の深い特徴量で

あると考えることができる。したがって、単語の出現回数を特徴量として、複数の文書を分類した場合、それぞれの分類カテゴリには意味の近い文書が所属すると考えることができる。

【０１３３】

選択部４０３により選択された一つあるいは複数の項目値は、特徴抽出部４０４に含まれる解析部４０９よって項目値ごとに形態素解析等の自然言語解析をおこない、単語に分割される。また、それぞれの単語には、その単語の品詞情報も付与される。出現した単語のうち、名詞であるものに対して一意な単語ＩＤを付与し、一つの項目値および選択部４０３により選択されたすべての項目値に対する単語ＩＤごとの出現回数を計数する。

【０１３４】

特徴抽出部４０４に含まれる特徴ベクトル生成部４１０は、計数された出現回数に基づいて個々の項目値の特徴（量）を示す項目値特徴ベクトルを生成する。たとえば、選択部４０３により選択された項目値が、

「騒音が大きい」

「塗装が変色する」

「オーバーヒートが起こる」

「塗装がはげる」

「バッテリーが上がる」

「排気が黒い」

であった場合、各項目の特徴ベクトルは、図１３に示すようになる。また、図１４には、単語とその単語ＩＤごとの出現回数を示す。

【０１３５】

すなわち、

「騒音が大きい」　　：{1，1，1，0，0，0，0，0，0，0，0，0，0}

「塗装が変色する」　　：{0，1，0，1，1，1，0，0，0，0，0，0，0}

「オーバーヒートが起こる」：{0，1，0，0，0，0，1，1，0，0，0

，０，０｝

「塗装がはげる」　　：｛０，１，０，１，０，０，０，０，１，０，０，０，０｝

「バッテリーが上がる」　：｛０，１，０，０，０，０，０，０，０，１，１，０，０｝

「排気が黒い」　　：｛０，１，０，０，０，０，０，０，０，０，０，１，１｝

という特徴ベクトルが得られる。

【０１３６】

この項目値の特徴ベクトルは、特徴抽出部４０４からの出力として加工処理部４０５へ送られる。加工処理部４０５においては、項目値の特徴ベクトルを用いて、分類処理をおこなう。分類処理は、まず、おのおののベクトル間の距離を計算することにより求める。距離の尺度には、たとえば内積を用いることができる。

【０１３７】

距離を計算した後、距離の近いものをまとめる処理をおこなう。この処理には、たとえばベクトルの集合をその距離に応じてＫ個のベクトルの集合に分類するＫ－ｍｅａｎｓ法を用いることができる。それぞれのベクトルの分類が完了したら、ベクトルに対応する項目値に対してどの分類に属するかの番号、すなわち、クラスタ番号と、項目値に対応する文書ＩＤを付与し、加工処理部４０５の出力とし、出力部４０６により出力表示をおこなう。

【０１３８】

図１５は、クラスタ番号１５０１を表示した画面の表示例を示す。クラスタ番号が同一番号である文書、たとえば、番号「１」および「６」の文書のクラスタ番号が「５」であり、両者が同一の分類に属することを示している。

【０１３９】

つぎに、請求項２の発明においてクロス表を出力する場合について説明する。入力部４０１により、分析対等とする文書群を読み込んだ後、操作者は分類処理をおこなう対象となる項目名、クロス表の横軸あるいは縦軸となる項目名、いく

つに分類をおこなうかの分類数を指示する。

【０１４０】

図１６はクロス表作成のための指示画面である。図１６において、指示画面１６００は、処理対象項目名入力欄１６０１と、軸となる項目名入力欄１６０２と、縦軸指定ボタン１６０３と、横軸指定ボタン１６０４と、分類数入力欄１６０５とから構成される。

【０１４１】

処理対象項目名入力欄１６０１には、処理対象となる項目名を入力する。キーボード２０９等から入力するあるいは処理対象となる項目候補を表示させその中からマウス２１０等により選択することにより項目名を入力することができる。また、軸となる項目名入力欄１６０２には、軸となる項目名を入力する。入力の方法は、処理対象項目名入力欄１６０１への入力の方法と同様である。

【０１４２】

縦軸指定ボタン１６０３および横軸指定ボタン１６０４は、軸となる項目を縦軸に表示させるか横軸に表示させるかを指定するためのボタンである。また、分類数入力欄１６０５には、いくつに分類するかその分類数を入力する。入力の方法としては、キーボード２０９等から数字を入力するあるいは分類数候補を表示させその中からマウス２１０等により選択することにより分類数を入力するようにしてもよい。

【０１４３】

図１６においては、処理対象項目名入力欄１６０１には「内容」が、軸となる項目名入力欄１６０２には「車種」が、また、横軸指定ボタン１６０４がチェックされ、分類数入力欄１６０５には「５０」が入力され、これにより、文書群の中の「内容」に基づいて、「５０（個）」に分類され、クロス表の横軸に「車種」を表示するという指示がなされていることがわかる。

【０１４４】

クロス表作成の指示がおこなわれることにより、分類処理が実行され、その結果がクロス表で表示される。図１７および図１８は、分類処理の結果が表示されたクロス表を示す図である。図１７において、クロス表１７００は、縦軸に分類

を示す「クラスタ１」、「クラスタ２」．．．が表示され、横軸に車種を示す「ＡＢＣ１６００」、「ＡＢＣ１８００」．．．が表示される。

【０１４５】

表の縦軸、すなわち各行は、分類処理により生成されたクラスタに対応する。各行の第１欄には、分類処理終了時には既定値としてクラスタ番号を示す文字列が入っている。表の横軸、すなわち各欄には、文書群の項目「車種」に含まれる文字列が重複することなく表示される。行「クラスタ１」の各セルには、クラスタ１に分類された文書のうち、項目「車種」の値がその欄の車種と一致するものの数が表示される。

【０１４６】

ここで、数を表示する代わりに、セルの色の濃淡や、セルを塗りつぶす面積により数の大きさを表現するようにしてもよい。また、表の最右欄および最下欄には、該当する行、欄の合計が表示される。

【０１４７】

図１８において、クロス表１７００のあるセルにマウスポインタ１８００を移動させ、マウス２１０のマウスボタンを押下する、あるいはキーボード２０９のカーソルキー操作によりカーソルを移動させ、特定キーを押下すると、そのセルの近傍に内容表示画面１８０１が表示されることにより、該当する文書の項目「内容」が表示される。

【０１４８】

内容表示画面１８０１には、セル内のデータ数、表示項目、セル情報、および、各データにおける表示項目の内容が表示される。マウスポインタ１８００により指定されたセルにおいては、データ数：「４」、表示項目：「内容」、セル情報：「ＡＢＣ２０００－クラスタ１」、表示項目の内容として「内容」の４つの内容である「排気が黒い、排気が黒い、．．．」が表示される。これにより、マウスポインタを所望のセルに移動させてマウスボタンを押下するという簡易な操作により、セルの内容を認識することができる。

【０１４９】

また、内容表示画面１８０１に表示される項目は、設定操作により変更するこ

とが可能であり、すべての項目を表示させることもでき、また、項目を選択して表示させることもできる。

【０１５０】

各行の第１欄には、分類処理終了時には既定値としてクラスタ番号を示す文字列が入っているが、操作者により、この欄の書き換えをすることができる。たとえば、上記の操作によってセルの内容を確認した後、「クラスタ１」を「排気の問題」と書き換えることができる。これにより、情報内容の把握がより容易になる。

【０１５１】

また、分類終了時に既定値としてクラスタ番号を示す文字列を入れるのではなく、そのクラスタの特徴を示す文字列を抽出し、セルに入れることも可能である。たとえば、クラスタ１に含まれる文書の項目「内容」から、もっとも頻度が高く出現する文や単語を抽出することにより実現する。

【０１５２】

図１８においては、クラスタ１には「排気が黒い」あるいは「排気」等の単語が入れられる。このように、操作者は簡易な操作により文書全体の分布状態を把握するだけでなく、必要に応じて個々の文書の内容をも詳細に知ることができる。

【０１５３】

つぎに、クロス表を作成するための出力部４０６の詳細な構成の内容について説明する。図１９は、実施の形態１による文書処理装置の出力部４０６の詳細な構成を示すブロック図である。出力部４０６は、グラフ描画部４０７のほかに、項目値選定部１９０１、集計部１９０２とから構成され、集計部はさらに実際に表示する内容に対応した記憶領域を持つ表保持部１９０３を備えている。

【０１５４】

項目値選定部１９０１は、操作者がクロス表の一つの軸として指定した項目名（軸項目名）に対して、文書記憶部４０２に記憶された文書データから、項目値を順次読み出し、重複のない項目値の集合を作成する。また、集計部１９０２は、表保持部１９０３の項目値に対応する領域に数値を加算することにより文書の

集計をおこなう。

【０１５５】

つぎに、クロス表の出力手順について説明する。図２０は、実施の形態１による文書処理装置のクロス表の出力手順を示すフローチャートである。図２０のフローチャートにおいて、まず、集計に先立ち、表保持部１９０３の内容を初期化する（ステップＳ２００１）。

【０１５６】

つぎに、項目値設定部１９０１により作成された項目値を、表の項目値ラベルに相当する部分に当てはめ（ステップＳ２００２）、クラスタ番号を表す文字列を、クラスタ番号に相当する部分に当てはめる（ステップＳ２００３）。

【０１５７】

つぎに、加工処理結果保持部４０８に保持された項目値に対応する文書ＩＤについて、文書記憶部４０２に記憶された対応する文書を参照し、その軸項目名に対応する項目値を決定する（ステップＳ２００４）。その後、表保持部１９０３の対応する領域の内容に１を加算する（ステップＳ２００５）。

【０１５８】

すべての項目値について上記処理をおこなったか否かを判断し（ステップＳ２００６）、すべての項目値について上記処理がおこなわれていない場合（ステップＳ２００６否定）は、ステップＳ２００４へ移行し、ステップＳ２００４～Ｓ２００６の処理を繰り返しおこなう。

【０１５９】

ステップＳ２００６において、すべての項目値について上記処理がおこなわれた場合（ステップＳ２００６肯定）は、最右列に表示するための行の合計を計算し（ステップＳ２００７）、併せて、最下行に表示するための欄の合計を計算する（ステップＳ２００８）。

【０１６０】

その後、表保持部１９０３に構成された表を、順次読み出して出力し（ステップＳ２００９）、すべての処理を終了する。

【０１６１】

なお、加工処理部４０５から出力されたデータを、文書記憶部４０２に送り、文書記憶部４０２に他のデータとともに記憶するように構成してもよい。文書記憶部４０２に記憶された加工処理部４０５から出力されたデータは、出力部４０６によって表の新たな列として表示することができる。また、表の既存の列を消去し、消去した列へ上書きするようにしてもよい。

【０１６２】

この構成では、処理の結果である加工処理部４０５から出力されたデータは、文書記憶部４０２において、今回の加工処理の対象とならなかった他のデータと対等に取り扱うことができ、その後の分析作業等で、もともとの入力データに存在していたか、分析作業の途中で加工処理によって生成されたのかを区別することなく、加工処理の対象として選択することが可能である。

【０１６３】

したがって、データの性質や、おこないたい情報分析作業の内容に応じて柔軟に加工処理対象と加工処理内容を選択することができるので、多彩で高度な情報分析作業が可能となる。

【０１６４】

また、加工処理部４０５への入力データとして、特徴抽出部４０４から出力されたデータだけではなく、選択部４０３により選択されたデータも含めることができる。これにより、文字列の特徴抽出を必要としないデータや、加工処理結果の数値に対してもさらなる加工処理を施すことができるので、より多彩で高度な情報分析が可能となる。

【０１６５】

図２１～図２４は、実施の形態１による文書処理装置の出力部４０６による画面表示の別の例を示す説明図である。図２１において、「番号」、「受付日」、「営業所」、「車種」、「年式」、「内容」の他に、分類処理により得られた結果である「クラスタ番号」２１０１が表示されている。

【０１６６】

さらに、図２１においては、選択部４０３により「クラスタ番号」２１０１が選択されており、「クラスタ番号」２１０１に関するデータが反転表示されてい

る。選択された「クラスタ番号」２１０１をキーとして、加工処理部４０５により並べ替え処理をおこなうよう指示をする。

【０１６７】

並べ替え処理の指示により、並べ替え処理がおこなわれた結果を表示しているのが図２２である。図２２においては、「クラスタ番号」が「１」のものが集まって表示されるように並べ替えられ、それに続き、「クラスタ番号」が「２」のものが集まって表示されるように並び替えられる。

【０１６８】

具体的には、「クラスタ番号」が「１」である「番号」が「２」、「１１」、「１５」、「２３」、「３５」、「５４」、「６３」、「７３」、「８２」の順で並べ替えられ、それに続き「クラスタ番号」が「２」である「番号」が「１４」、「１８」、「２２」、「２７」、「３７」、．．．が表示されていることがわかる。

【０１６９】

つぎに、項目「車種」の欄で、「クラスタ番号」が「１」に属するものを選択する。図２３においては、項目「車種」の欄で、「クラスタ番号」が「１」に属するものが選択され、その選択領域２３０１が反転表示されていることを示している。このように、すでに「クラスタ番号」により並べ替えがおこなわれており、同一クラスタに属するものが集まって表示されているので、画面上の連続した領域として容易に選択することができる。

【０１７０】

つぎに、選択領域２３０１について車種別の発生頻度の棒グラフを表示させたのが、図２４である。図２４において、棒グラフ表示領域２４０１には、選択領域２３０１によって選択された「クラスタ番号」が「１」である９つの文書が選択され、その９つの文書を車種別に棒グラフ化したものが表示される。

【０１７１】

このように、加工処理の対象を柔軟かつ容易に選択でき、選択された対象について多様な加工処理をおこなうことができ、また、その加工処理結果も次回の加工処理の対象とすることができるので、高度な情報分析作業が可能となる。

【０１７２】

このように、分類等の文字列の特徴量を抽出して、その特徴量を用いておこなう加工処理を実施した後に多種の加工処理をおこなう例を示したが、事前に多種の処理をおこなうことができるようにしてもよい。

【０１７３】

たとえば、「車種」の項目を選択し、これをキーとして並べ替えをおこなった後、集まったある車種、たとえば、「ＡＢＣ１６００」に対して分類処理をおこなうこともできる。また、入力部４０１により入力された文書が誤字等の誤りを含んでいる場合、分類等の文字列の特徴量を抽出して、その特徴量を用いて加工処理をおこなう前に、たとえば、文字列の検索・置換処理をおこなって、誤字を一括して修正し、より好適な結果が得られるようにデータを整えることもできる。

【０１７４】

図２５は、実施の形態１による文書処理装置の文書記憶部４０２の詳細な構成を示すブロック図である。図２５において、文書記憶部４０２は、設定値記憶部２５０１および設定値送受信部２５０２を含んでいる。設定値記憶部２５０１には、文書を分類する際の分類数等の分類情報記憶部２５０３をはじめとするさまざな設定値、すなわち文書処理装置の動作に必要な設定値に関する情報を記憶する記憶部を備えている。これにより設定値に関する情報は、文書情報とともに記憶することができる。

【０１７５】

また、設定値送受信部２５０２は、設定値記憶部２５０１によって記憶された設定値に関する情報を他の情報処理装置へ送信する。また、設定値送受信部２５０２は、他の情報処理装置からの設定値に関する情報を受信する。設定値送受信部２５０２により受信された設定値に関する情報は、設定値記憶部２５０１によって記憶される。

【０１７６】

記憶された設定値に関する情報は、後に文書を再度読み込んだときに同時に読み込まれ設定値記憶部２５０１に記憶される。この設定値に関する情報は操作者

が所定の操作をすることにより参照することができたり、以後の処理の際に、再利用することができる。これにより、設定値に関する情報を文書とともに保存・管理することが可能となるので、設定値に関する情報の紛失を防ぎ、好適な設定値を後に再利用することができる。

【０１７７】

図２６～図２８は、実施の形態１による文書処理装置の出力部４０６による画面表示の別の例を示す説明図である。図２６において、まず、操作者が分類をおこなうべき対象である「内容」を表示画面上で選択する。それにより選択領域２６０１が反転表示される。つぎに、メニュー・バー２６０３から、分類処理ボタン２６０３を選択すると、分類処理に必要な分類数、すなわち、対象をいくつに分類するかについての問い合わせ画面２６０４が表示される。

【０１７８】

操作者が問い合わせ画面２６０４において分類数を入力すると、この分類数に関する情報が文書記憶部４０２に記憶される。図２６においては、分類数として「５０」が入力されたことを示している。

【０１７９】

その後、操作者が情報分析作業を完了して、メニュー・バー２６０３のファイルボタン２６０５の選択によりポップアップする図示を省略する保存ボタンを押下すると、文書記憶部４０２により、操作者が指示したファイル名が付与され、文書の情報、分類結果とともに記憶される。

【０１８０】

図２７において、分類結果を表示する欄２７０１にマウスポインタ２７０２を移動させ、マウスボタンを押下すると、その分類をおこなうことに用いた分類に関する情報および分類設定値に関する情報を表示する分類情報表示画面２７０３が表示される。これにより、用いた設定値の関連づけが容易に把握することができる。

【０１８１】

分類情報表示画面２７０３には、たとえば、分類に関する情報として分類がおこなわれた日時に関する情報を示す「分類日時」、分類の対象となった文書数に

関する情報を示す「分類対象数」等が表示され、また、分類設定値に関する情報として、いくつに分類したかを示す「分類数」、どの品詞に基づいて分類をしたかを示す「分類品詞」等が表示される。

【０１８２】

分類処理を実行するたびに新規な表が作成される。図２８は、分類結果１を得た後、再度分類処理がおこなわれ、分類結果２が表示された状態を示している。分類結果１を再度表示させたい場合は、画面左下部のラベル上の選択領域２８０１へマウスポインタを移動させ、マウスボタンを押下する。これにより、分類結果１が再度表示される。その後、分類結果２を再度表示させる場合も同様の操作によりおこなうことができる。

【０１８３】

また、図２８において、各分類処理の実行に用いた設定値に関する情報が対応する表の所定の表示領域２８０２に表示される。この表示領域２８０２は、分類結果の表示を隠さないように表示させることができ、また、その表示位置を移動することもできる。これにより、分類結果と、それに用いた設定値の関連づけが容易に把握できる。

【０１８４】

つぎに、実施の形態１における文書処理装置の文書処理の一連の手順について説明する。図２９は、実施の形態１による文書処理装置の文書処理の一連の手順を示すフローチャートである。

【０１８５】

図２９のフローチャートにおいて、まず、文書データが文書処理装置に入力されたか否かを判断する（ステップＳ２９０１）。ここで、文書データが入力されるのを待って、文書データが入力された場合（ステップＳ２９０１肯定）は、入力された文書データを記憶する（ステップＳ２９０２）。なお、ステップＳ２９０１およびＳ２９０２の各ステップは、文書の入力があるごとに他のステップとは独自におこなわれるようにしてもよい。

【０１８６】

つぎに、記憶された文書データの全部または一部が選択されたか否かを判断す

る（ステップＳ２９０３）。ここで、文書データの全部または一部が選択されるのを待って、選択された場合（ステップＳ２９０３肯定）は、選択された文書データの全部または一部の文字列の特徴に関するデータの抽出をおこなう（ステップＳ２９０４）。

【０１８７】

その後、ステップＳ２９０４において、抽出された文字列の特徴に関するデータに基づいて、分類処理等、所定の加工処理をおこなう（ステップＳ２９０５）。続いて、ステップＳ２９０５において加工処理がおこなわれたデータを、表形式に展開する等の出力処理をおこなう（ステップＳ２９０６）。

【０１８８】

さらに、ステップＳ２９０５において加工処理されてデータを元の文書データに関連づけして記憶する（ステップＳ２９０７）。また、加工処理の設定値等の加工処理の内容に関するデータも併せて記憶する（ステップＳ２９０８）。

【０１８９】

その後、ステップＳ２９０５において加工処理されたデータの全部または一部が選択されたか否かを判断し（ステップＳ２９０８）、選択されなかった場合（ステップＳ２９０９否定）は、ステップＳ２９０４へ移行し、以後、ステップＳ２９０４～Ｓ２９０９の処理を繰り返しおこなう。一方、ステップＳ２９０９において、加工処理されたデータの全部または一部が選択された場合（ステップＳ２９０９肯定）は、すべての処理を終了する。

【０１９０】

なお、実施の形態１で説明した文書処理方法は、あらかじめ用意されたプログラムをパーソナルコンピュータやワークステーション等のコンピュータで実行することにより実現される。このプログラムは、ハードディスク、フロッピーディスク、ＣＤ－ＲＯＭ、ＭＯ、ＤＶＤ等のコンピュータで読み取り可能な記録媒体に記録され、コンピュータによって記録媒体から読み出されることによって実行される。またこのプログラムは、上記記録媒体を介して、または伝送媒体として、インターネット等のネットワークを介して配布することができる。

【０１９１】

つぎに、実施の形態２～６に係る情報分類装置について説明する。なお、以下説明する実施の形態２～６においては、上記のように多くのノイズを含んだものであるとの解釈に基づいて、一回の文書集合からの話題（内容）抽出と位置づけ、文書分類のためのパラメータ（対象文書集合やクラスタ数、類似度測度、ストップワード等）を変化させながら複数化の分類を実行させ、その結果を保持・統合する手段を設けることで、任意の文書集合にどのような内容が含まれるかを漸次的に収集するものである。

【０１９２】

〔実施の形態２〕

この発明の実施の形態２に係る文書分類装置を構成する情報処理システムは、図１に示したように実施の形態１の情報処理システムと同様であるので、その説明は省略する。また、サーバー１０１およびクライアント１０２のハードウエア構成についても、図２・図３に示したように実施の形態１と同様であるので、その説明は省略する。

【０１９３】

つぎに、実施の形態２による文書分類装置の機能的構成について説明する。図３０は、実施の形態２による文書分類装置の構成を機能的に示すブロック図である。

【０１９４】

図３０のブロック図において、文書分類装置は、入力部３００１と、言語解析部３００２と、ベクトル生成部３００３と、分類部３００４と、分類パラメータ指示部３００５と、分類結果記憶部３００６と、クラスタ特徴表示部３００７と、クラスタ特徴算出部３００８と、分類体系記憶部３００９と、クラスタ選択指示部３０１０と、分類体系閲覧操作部３０１１と、を含む構成である。

【０１９５】

入力部３００１、言語解析部３００２、ベクトル生成部３００３、分類部３００４、分類パラメータ指示部３００５、分類結果記憶部３００６、クラスタ特徴表示部３００７、クラスタ特徴算出部３００８、分類体系記憶部３００９、クラスタ選択指示部３０１０、分類体系閲覧操作部３０１１は、ＲＯＭ２０２または

３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０１９６】

ここで、入力部３００１は、文書データを入力するものであり、たとえば、キーボード２０９または３１１、スキャナ３１３、ＯＣＲ機能を備えたスキャナ３１３、またはネットワーク１０３を経由して文書や文書群を得ることができるＩ／Ｆ２０４または３０９等である。

【０１９７】

また、入力部３００１は、上記以外に、文書データを取得することができるものであれば、それらのすべてを含む。たとえば、文書データがデータベース化されている場合に、そのデータベースが記録された媒体を本実施の形態の文書分類装置に組み入れた場合も文書データの入力とする。

【０１９８】

また、言語解析部３００２は、入力部３００１により入力された文書データを解析して言語解析情報を得るものであり、ベクトル生成部３００３は、言語解析部３００２により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成するものである。

【０１９９】

また、分類部３００４は、ベクトル生成部３００３により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成するものであり、分類パラメータ指示部３００５は、分類パラメータを指示するものであり、たとえば、キーボード２０９または３１１、マウス２１０または３１２、またはネットワーク１０３を経由して指示情報を得ることができるＩ／Ｆ２０４または３０９等である。

【０２００】

また、分類結果記憶部３００６は、分類部３００４により分類された結果、すなわち、分類された文書の部分集合に関する情報を記憶するものである。また、

クラスタ特徴表示部３００７は、クラスタ特徴算出部３００８により算出されたクラスタ特徴を表示する。

【０２０１】

クラスタ特徴算出部３００８は、分類部３００４により生成された文書の部分集合の特徴であるクラスタ特徴を算出するものである。また、分類体系記憶部３００９は、クラスタ特徴算出部３００８により算出されたクラスタ特徴を分類体系の構成要素として記憶するものである。また、分類体系記憶部３００９は、クラスタ選択指示部３０１０により選択された文書の部分集合を分類体系の構成要素として記憶するものである。すなわち、クラスタ選択指示部３０１０により選択されたクラスタに所属する全ての文書もしくは所属する文書の一部を分類体系の構成要素として記憶するものである。

【０２０２】

クラスタ選択指示部３０１０は、クラスタ表示部３００７により表示された複数のクラスタ特徴の中から所望のクラスタを選択するものである。また、クラスタ選択指示部３０１０は、前記分類部３００４により生成された文書の部分集合の中から所望の部分集合を選択するものである。また、分類体系閲覧操作部３０１１は、分類体系記憶部３００９に記憶されたデータを閲覧したい場合に、その閲覧の操作をおこなうものである。

【０２０３】

つぎに、文書集合に含まれる話題（内容）を抽出することが重要となる好適な例を、アンケート調査等により得られた自由記述回答の分析場面を想定し、その具体例を用いて説明する。

【０２０４】

近年、たとえば、インターネット等を介して短期間に数千～数万件の自由記述回答を回収することが可能であり、このような機能を用いて大量のテキスト情報の収集をおこなうことができる。

【０２０５】

アンケート調査により得られた大量のテキスト情報の収集の例として、「オフィスのネットワーク化による無駄を挙げてください」という質問に対して文書で

答えた一つの回答記述を文書とすると、文書集合（クラスタ）は１件ごとの回答の集合ということになる。

【０２０６】

ここで、操作者（アンケートの分析者）は、そのニーズの一つとして、意見集合（文書集合）にどのような種類の意見（話題）が含まれており、意見の概略を把握したい場合がある。このようなニーズを満たすべく、話題の抽出を類似する意見のまとまり（分類）により実現し、アンケート結果にどのような種類の意見が含まれているかを抽出する。

【０２０７】

文書分類は、典型的には大きく分けてつぎの３段階のステップから構成される。第１ステップでは、入力部３００１により入力された各文書（意見）について、言語解析部３００２が、各文書に含まれる単語（あるいは、特定の連続する文字列）を抽出する。この際、たとえば、形態素形跡等の言語解析アルゴリズムが用いられる。

【０２０８】

第２ステップでは、抽出された単語を列とし、各文書を行とし、要素を単語の出現頻度とした「単語」×「文書」の行列が生成される。なお、一般的な形態素解析機能と構文解析機能を有する言語解析ツールを用いると単語抽出のほかに、単語の品詞情報、複合語（フレーズ）、構文情報等の同時に取得することができ、こうした情報を上記単語×文書の行列を生成する際、考慮することができる。

【０２０９】

ベクトル生成部３００３は、この「単語」×「文書」の行列に基づいて単語で構成される多次元空間内に各文書をベクトル表現する。これには、以下の方法があり、本実施の形態においては、すべての方法を実装している。

【０２１０】

（１）行列の列成分をそのまま利用する方法、

（２）各文書の長さ（文字の数やページ数等）や分類対象全体の文書集合内での各単語の出現頻度を考慮して値の重み付けをする方法、

（３）上記行列から文書間の内積行列を算出し、これに特異値分解（たとえば、

因子分析や主成分分析、数量化理論第３類等を利用しておこなわれる）を適用して潜在的意味空間を構成する方法、
等である。

【０２１１】

また、「Ｒｅｐｒｅｓｅｎｔａｔｉｎｇ Ｄｏｃｕｍｅｎｔｓ Ｕｓｉｎｇ ａｎ Ｅｘｐｌｉｃｉｔ Ｍｏｄｅｌ ｏｆ Ｔｈｅｉｒ Ｓｉｍｉｌａｒｉｔｉｅｓ（著者名：Ｂｒｉａｎ Ｔ．Ｂａｒｔｅｌｌ，Ｇａｒｒｉｓｏｎ Ｗ．Ｃｏｔｔｒｅｌｌ，ａｎｄ Ｒｉｃｈａｒｄ Ｋ．Ｂｅｌｅｗ，論文名：Ｊｏｕｒｎａｌ ｏｆ ｔｈｅ Ａｍｅｒｉｃａｎ Ｓｏｃｉｅｔｙ ｆｏｒ Ｉｎｆｏｒｍａｔｉｏｎ Ｓｃｉｅｎｃｅ，学会名：ｔｈｅ Ａｍｅｒｉｃａｎ Ｓｏｃｉｅｔｙ ｆｏｒ Ｉｎｆｏｒｍａｔｉｏｎ Ｓｃｉｅｎｃｅ，ページ：２５４－２７１，Ｖｏｌ．４６ Ｎｏ．４，発行年：１９９５）」においては、上記潜在的意味空間への変換手法を一般化し、文書間の内積行列に、文書が有するほかの文書への参照情報から生成される共参照情報などを付加した行列を用いて、これらの類似性を反映する空間へ文書や単語を射影するための表現空間変換関数を導出しているものもあり、この方法も利用することができる。

【０２１２】

第３ステップでは、分類部３００４が、文書特徴ベクトルの類似度を用いて文書を分類する。具体的には分類対象データに対してカイ自乗法の手法、判別分析の方法、クラスタリングの方法等を適用することにより分類が実行される。

【０２１３】

また、類似度としては、内積や余弦、ユークリッド距離、マハラノビスの距離等が考えられ、本実施の形態においては、いずれの方法を用いてもよい。

【０２１４】

また、クラスタリングのアルゴリズムに関してもさまざまなものが公知になっている。クラスタリングは、大別して階層型クラスタリングと非階層型クラスタリングが考えられるが、本実施の形態においては、いずれの方法を用いてもよい。

【０２１５】

また、分類パラメータ指示部３００５は、分類部３００４が文書特徴ベクトルを分類するための分類パラメータを指示する。分類部３００４は、分類パラメータ指示部３００５により指示された分類パラメータにしたがって内部に保持される文書特徴ベクトルを分類する。

【０２１６】

このようにして、第１ステップ～第３ステップの各処理を実行することにより第１回目の文書分類が終了すると、分類結果は分類結果記憶部３００６により保持される。

【０２１７】

引き続き、クラスタ特徴算出部３００８が、分類結果がどのようなクラスタを得ることができたのかを示す特徴、すなわちクラス特徴を算出する。典型的には各クラスタに所属する文書、あるいはその文書の一部を算出するが、その際、クラスタの重心との類似度に基づいて文書をソーティングして出力する。

【０２１８】

そのほか、クラスタ内で最頻の単語、クラスタに所属する文書数、クラスタ内での文書のばらつきの程度を表すクラスタ内の標準偏差のような数値をクラスタの特徴を表現するものとして算出する。

【０２１９】

これらのクラスタの特徴情報は、操作者に対して出力（表示）されたクラスタがどのようなもの（どのような特徴を有するもの）かを把握させるために算出されるものであり、操作者に対してクラスタの特徴を示すものであれば、上記の内容（特徴）以外のものであってもよい。

【０２２０】

また、クラスタ特徴算出部３００８は、上記のようにクラスタの特徴を示すもの以外に、クラスタ間の関係を示す情報も算出する。階層型クラスタリングの場合は、その上位あるいは下位のクラスタを、非階層型クラスタリングの場合は、クラスタ重心間の類似度に基づく近接のクラスタを算出する。

【０２２１】

つぎに、クラスタ特徴表示部３００７によるクラスタ特徴の表示およびクラスタ選択の内容について説明する。図３１は、実施の形態２による文書分類装置のクラスタ特徴表示部３００７の表示の一例を示す説明図である。

【０２２２】

図３１において、クラスタ単位で操作者ができるようになっており、各クラスタは「クラスタＩＤ」欄３１０１、「メンバー数」欄３１０２、「頻度の高い単語」欄３１０３、「文書内容」欄３１０４、「重心との類似度」欄３１０５等の項目から構成される。

【０２２３】

「クラスタＩＤ」欄３１０１には、クラスタのＩＤを示す番号が通し番号で付与され、表示される。「メンバー数」欄３１０２はクラスタに所属する文書あるいは文書の一部の数が算出され、表示される。その中で頻度の高い単語が抽出され「頻度の高い単語」欄３１０３に表示される。「文書内容」欄３１０４には文書の内容が表示され、「重心との類似度」欄３１０５には、数値化された重心との類似度が表示される。これにより、操作者の理解容易性が向上する。

【０２２４】

操作者は、表示された情報（特徴量）に基づいてクラスタについてその特徴を把握することができる。ここで、内容（特徴）が理解可能なクラスタが一つでもあれば、操作者はクラスタ選択指示部３０１０によりクラスタを選択することができる。

【０２２５】

より具体的には、マウス２１０または３１２等によって、表示されているクラスタの所定の位置、たとえば、「クラスタＩＤ」欄３１０１へカーソル３１１０を移動させ、その位置でクリックすることにより、当該クラスタＩＤのクラスタ全体を選択することができる。なお、選択したクラスタに所属する文書は必ずすべてが選択されるわけではなく、その一部の文書が選択されるようにしてもよい。

【０２２６】

図３１においては、「クラスタＩＤ」欄３１０１がクリックされ、これにより

、クラスタ全体が反転表示しており、当該クラスタ（クラスタＩＤ「1」）が選択されたことを示している。

【０２２７】

また、操作者は、内容が理解可能であるクラスタが存在しない場合は、分類パラメータ指示部３００５により分類パラメータの再設定をおこない、再度分類実行をおこなうことができる。

【０２２８】

クラスタ選択指示部３０１０により選択されたクラスタＩＤに関するデータは分類体系記憶部３００９へ送信される。分類体系記憶部３００９は、このクラスタＩＤに関するデータに基づいてクラスタ特徴算出部３００８からクラスタに関する上記特徴量を検索し記憶する。

【０２２９】

また、分類体系記憶部３００９は、同様に、分類結果記憶部３００６から分類結果を検索し記憶する。さらに、分類体系記憶部３００９は、操作者により入力されたクラスタに関するコメント（たとえば、「ネットワークの維持費が高い」等）の情報を併せて記憶することもできる。このように、操作者が作成した情報を分類体系の構成要素として記憶することにより、分類体系の利用価値がより向上する。

【０２３０】

なお、分類体系記憶部３００９により記憶されたデータは、別途閲覧操作用のインターフェイスを設けることにより、選択・保持するクラスタの内容の閲覧や、クラスタ間の意味的な関連を手動であるいは、保持されているクラスタ重心間の類似度等を用いて自動で、構造化・体系化することができる。

【０２３１】

つぎに、実施の形態２の文書分類装置の一連の処理の手順について説明する。図３２は、実施の形態２による文書分類装置の一連の処理の手順を示すフローチャートである。図３２のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップＳ３２０１）。

【０２３２】

つぎに、入力された文書の言語が解析され（ステップＳ３２０２）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成される（ステップＳ３２０３）。

【０２３３】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップＳ３２０４肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップＳ３２０５）、その結果、すなわち、クラスタに関する情報を記憶する（ステップＳ３２０６）。

【０２３４】

つぎに、分類されたクラスタの特徴を算出し（ステップＳ３２０７）、算出された結果を表示する（ステップＳ３２０８）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップＳ３２０９）、選択されなかった場合（ステップＳ３２０９否定）は、ステップＳ３２０４へ移行し、再度分類パラメータの指示があるのを待つ（ステップＳ３２０４）。

【０２３５】

一方、ステップＳ３２０９において、クラスタが選択された場合（ステップＳ３２０９肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップＳ３２１０）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。これにより、一連の処理を終了する。

【０２３６】

以上説明したように、実施の形態２による文書分類装置によれば、分類対象である文書群での文書間の類似性に基づいて、各文書をそれら文書間の意味的な関連性を反映しうる表現空間へ変換するための表現空間変換関数を算出し、その表現空間で文書分類をおこなうことにより、操作者の意図を反映しうる文書分類を実現することができる。

【０２３７】

したがって、分類部３００４によりクラスタを得ることができるとともに、クラスタ特徴算出部３００８・分類体系記憶部３００９により、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこな

うことができる。

【０２３８】

また、クラスタ選択指示部３０１０により選択されたクラスタのみを用いて、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができる。

【０２３９】

〔実施の形態３〕

さて、上述した実施の形態２に加えて、以下に説明する実施の形態３のように、さらにベクトル記憶部と、ベクトル修正部とを含む構成とするようにしてもよい。

【０２４０】

実施の形態３による文書分類装置を構成する情報処理システムは、図１に示したように実施の形態１と同様であるので、その説明は省略する。また、サーバー１０１およびクライアント１０２のハードウエア構成についても、図２・図３に示したように実施の形態１と同様であるので、その説明は省略する。

【０２４１】

つぎに、実施の形態３による文書分類装置の機能的構成について説明する。図３３は、この発明の実施の形態３による文書分類装置の構成を機能的に示すブロック図である。図３３において、実施の形態２の図３０と同一のものに関しては同じ符号を付して、その説明を省略する。

【０２４２】

図３３のブロック図において、文書分類装置は、入力部３００１、言語解析部３００２、ベクトル生成部３００３、分類部３００４、分類パラメータ指示部３００５、分類結果記憶部３００６、クラスタ特徴表示部３００７、クラスタ特徴算出部３００８、分類体系記憶部３００９、クラスタ選択指示部３０１０、分類体系閲覧操作部３０１１のほか、ベクトル記憶部３３０１と、ベクトル修正部３３０２とを含む構成である。

【０２４３】

ベクトル記憶部３３０１は、ベクトル生成部３００３により生成された文書特

徴ベクトルを記憶するものである。また、ベクトル修正部３３０２は、文書特徴ベクトル記憶部３３０１により記憶された文書特徴ベクトルを、クラスタ選択指示部３０１０により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正するものである。

【０２４４】

また、分類部３００４は、ベクトル修正部３３０２により修正された文書特徴ベクトルに基づいて文書を分類する。

【０２４５】

なお、ベクトル記憶部３３０１、ベクトル修正部３３０２は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０２４６】

ベクトル生成部３００３において生成された文書特徴ベクトル（列ベクトル）・単語（単語特徴）ベクトル（行ベクトル）はベクトル記憶部３３０１によって記憶される。これは、次回以降の分類実行の際に利用する文書特徴ベクトルを確保するためである。

【０２４７】

ベクトル修正部３３０２は、クラスタ選択指示部３０１０により選択されたクラスタに所属する文書のすべてあるいはその一部の文書を除き、次回以降もこれらの文書が除かれるよう削除する。削除された文書特徴ベクトルはベクトル記憶部３３０１により記憶される。

【０２４８】

この結果、ベクトル記憶部３３０１に記憶されているベクトルデータのうち、選択されたクラスタに所属する文書（もしくは操作者に指定されたその一部）列ベクトルを除いたものが、次回以降の分類が実行される際に利用されるデータとなる。

【０２４９】

つぎに、実施の形態３の文書分類装置の一連の処理の手順について説明する。図３４は、実施の形態３よる文書分類装置の一連の処理の手順を示すフローチャートである。図２のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップＳ３４０１）。

【０２５０】

つぎに、入力された文書の言語が解析され（ステップＳ３４０２）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され（ステップＳ３４０３）、生成された文書特徴ベクトルが記憶される（ステップＳ３４０４）。

【０２５１】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップＳ３４０５肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップＳ３４０６）、その結果、すなわち、クラスタに関する情報を記憶する（ステップＳ３４０７）。

【０２５２】

つぎに、分類されたクラスタの特徴を算出し（ステップＳ３４０８）、算出された結果を表示する（ステップＳ３４０９）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップＳ３４１０）、選択されなかった場合（ステップＳ３４１０否定）は、ステップＳ３４０５へ移行し、再度分類パラメータの指示があるのを待つ（ステップＳ３４０５）。

【０２５３】

一方、ステップＳ３４１０において、クラスタが選択された場合（ステップＳ３４１０肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップ３４１１）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する（ステップＳ３４１２）。

【０２５４】

ステップＳ３４１２において、繰り返して処理をおこなう旨の指示があった場合（ステップＳ３４１２肯定）は、選択されたクラスタに所属する文書のすべて

あるいはその一部の文書を除くように文書特徴ベクトルを修正する（ステップＳ３４１３）。その後、ステップＳ３４０５へ移行し、以後、ステップＳ３４０５～Ｓ３４１３の各処理を繰り返しおこなう。

【０２５５】

一方、ステップＳ３４１２において、繰り返して処理をおこなう旨の指示がない場合（ステップＳ３４１２否定）は、これにより、一連の処理をすべて終了する。

【０２５６】

以上説明したように、実施の形態３による文書分類装置によれば、ベクトル修正部３３０１により、既知になったクラスタの影響を排除した新たなクラスタを生成することができる。

【０２５７】

〔実施の形態４〕

さて、上述した実施の形態３においては、ベクトル記憶部およびベクトル修正部とを含む構成であったが、以下に説明する実施の形態４のように、ベクトル修正部に代わりに、文書表現空間修正部を含む構成とするようにしてもよい。

【０２５８】

実施の形態４による文書分類装置を構成する情報処理システムは、図１に示したように実施の形態１と同様であるので、その説明は省略する。また、サーバー１０１およびクライアント１０２のハードウエア構成についても、図２・図３に示したように実施の形態１と同様であるので、その説明は省略する。

【０２５９】

つぎに、実施の形態４による文書分類装置の機能的構成について説明する。図３５は、この発明の実施の形態４による文書分類装置の構成を機能的に示すブロック図である。図３５において、実施の形態２の図３０と同一のものに関しては同じ符号を付して、その説明を省略する。

【０２６０】

図３５のブロック図において、文書分類装置は、入力部３００１、言語解析部３００２、ベクトル生成部３００３、分類部３００４、分類パラメータ指示部３

００５、分類結果記憶部３００６、クラスタ特徴表示部３００７、クラスタ特徴算出部３００８、分類体系記憶部３００９、クラスタ選択指示部３０１０、分類体系閲覧操作部３０１１のほか、ベクトル記憶部３５０１と、文書表現空間修正部３５０２とを含む構成である。

【０２６１】

ベクトル記憶部３５０１は、ベクトル生成部３００３により生成された文書特徴ベクトルを記憶するものである。また、文書表現空間修正部３５０２は、文書特徴ベクトル記憶部３５０１により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示部３０１０により選択された部分集合から算出する特徴量に基づいて修正するものである。

【０２６２】

また、分類部３００４は、文書表現空間修正部３５０２により修正された文書表現空間を用いて、ベクトル生成部３００３により生成された文書特徴ベクトル間の類似度に基づいて文書を分類する。

【０２６３】

なお、ベクトル記憶部３５０１、文書表現空間修正部３５０２は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０２６４】

つぎに、文書表現空間修正部３５０２の内容について説明する。実施の形態３におけるベクトル修正部３３０２にあっては、既知になったクラスタの影響を排除するために文書特徴ベクトルを除去するが、文書特徴ベクトルを表現する多次元空間自体の変更はおこなわれない。

【０２６５】

したがって、前回の分類実行の結果、操作者により選択されたクラスタの形成特徴を次回の分類実行の際に排除したい場合は、文書ベクトルを表現する空間自体の変更が必要となる。

【０２６６】

そこで、文書表現空間修正部３５０２を備え、文書表現空間の修正をおこなうものである。ここで、文書表現空間の特徴次元を変更する例として、操作者により選択されたクラスタの重心と類似度の高い特徴次元の削除をおこなうことについて説明する。

【０２６７】

操作者により選択されたクラスタの重心はベクトルとして表現することができるので、このクラスタ重心ベクトルとベクトル記憶部３５０１に記憶されている文書表現空間の各特徴次元との類似度を算出することにより、類似度の高い特徴次元を判別する。

【０２６８】

なお、類似の測度としては、余弦、内積、ユークリッド距離、マハラノビス距離等を用いる。また、判別に関してはある類似度以上を削除対象として採用するようなしきい値処理による判別や、類似度の高い順にある一定数を削除対象として採用する定数処理による判別を用いる。また、判別分析等も用いることができる。

【０２６９】

文書表現空間修正部３５０２は、上述のような削除対象の特徴次元を算出して、特徴次元の削除をおこなう。特徴次元の削除は、ベクトル記憶部３５０１に記憶されている「特徴次元（単語）」×「文書」の行列から判別された特徴次元について行ベクトルを削除することによりおこなう。文書表現空間修正部３５０２により修正された文書ベクトルは、次回以降の分類のために、ベクトル記憶部３５０１に記憶される。

【０２７０】

つぎに、実施の形態４の文書分類装置の一連の処理の手順について説明する。図３６は、実施の形態４よる文書分類装置の一連の処理の手順を示すフローチャートである。図３６のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップＳ３６０１）。

【０２７１】

つぎに、入力された文書の言語が解析され（ステップＳ３６０２）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され（ステップＳ３６０３）、生成された文書特徴ベクトルが記憶される（ステップＳ３６０４）。

【０２７２】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップＳ３６０５肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップＳ３６０６）、その結果、すなわち、クラスタに関する情報を記憶する（ステップＳ３６０７）。

【０２７３】

つぎに、分類されたクラスタの特徴を算出し（ステップＳ３６０８）、算出された結果を表示する（ステップＳ３６０９）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップＳ３６１０）、選択されなかった場合（ステップＳ３６１０否定）は、ステップＳ３６０５へ移行し、再度分類パラメータの指示があるのを待つ（ステップＳ３６０５）。

【０２７４】

一方、ステップＳ３６１０において、クラスタが選択された場合（ステップＳ３６１０肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップ３６１１）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する（ステップＳ３６１２）。

【０２７５】

ステップＳ３６１２において、繰り返して処理をおこなう旨の指示があった場合（ステップＳ３６１２肯定）は、「特徴次元（単語）」×「文書」の行列から判別された特徴次元について行ベクトルを削除することにより文書表現空間を修正する（ステップＳ３６１３）。その後、ステップＳ３６０５へ移行し、以後、ステップＳ３６０５～Ｓ３６１３の各処理を繰り返しおこなう。

【０２７６】

一方、ステップＳ３６１２において、繰り返して処理をおこなう旨の指示がな

かった場合（ステップＳ３６１２否定）は、これにより、一連の処理を終了する。

【０２７７】

以上説明したように、実施の形態４による文書分類装置によれば、前回の分類実行の結果、文書表現空間修正部３５０２により操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【０２７８】

〔実施の形態５〕

さて、上述した実施の形態３または実施の形態４においては、ベクトル修正部または文書表現空間修正部のいずれか一方のみを含む構成であったが、以下に説明する実施の形態５のように、ベクトル修正部および文書表現空間修正部の両方を含む構成とするようにしてもよい。

【０２７９】

実施の形態５による文書分類装置を構成する情報処理システムは、図１に示したように実施の形態１と同様であるので、その説明は省略する。また、サーバー１０１およびクライアント１０２のハードウエア構成についても、図２・図３に示したように実施の形態１と同様であるので、その説明は省略する。

【０２８０】

つぎに、実施の形態５による文書分類装置の機能的構成について説明する。図３７は、この発明の実施の形態５による文書分類装置の構成を機能的に示すブロック図である。図３７において、実施の形態２の図３０と同一のものに関しては同じ符号を付して、その説明を省略する。

【０２８１】

図３７のブロック図において、文書分類装置は、入力部３００１、言語解析部３００２、ベクトル生成部３００３、分類部３００４、分類パラメータ指示部３００５、分類結果記憶部３００６、クラスタ特徴表示部３００７、クラスタ特徴算出部３００８、分類体系記憶部３００９、クラスタ選択指示部３０１０、分類体系閲覧操作部３０１１のほか、ベクトル記憶部３７０１と、ベクトル修正部３

７０２と、文書表現空間修正部３７０３とを含む構成である。

【０２８２】

ベクトル記憶部３７０１は、ベクトル生成部３００３により生成された文書特徴ベクトルを記憶するものである。また、ベクトル修正部３７０２は、文書特徴ベクトル記憶部３７０１により記憶された文書特徴ベクトルを分類部３００４により生成された文書の部分集合の文書特徴ベクトルを除去したのこりの文書特徴ベクトルとなるように修正するものである。

【０２８３】

また、文書表現空間修正部３７０３は、ベクトル記憶部３７０１により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示部３０１０により選択されたクラスタ特徴に基づいて修正するものである。

【０２８４】

また、分類部３００４は、文書表現空間修正部３７０３により修正された文書表現空間を用いて、ベクトル修正部３７０２により修正された文書特徴ベクトル間の類似度に基づいて文書を分類する。

【０２８５】

なお、ベクトル記憶部３７０１、ベクトル修正部３７０２、文書表現空間修正部３７０３は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現する。

【０２８６】

つぎに、ベクトル修正部３７０２および文書表現空間修正部３７０３の内容について説明する。実施の形態４においては、選択されたクラスタに所属する文書は次回以降の分類実行の際にも使用される。

【０２８７】

実施の形態５では、ベクトル修正部３７０２および文書表現空間修正部３７０３の両方を具備することにより、選択されたクラスタに所属する文書を次回の分

類実行の際に除去し、次回の分類実行の際には分類対象文書としないようにする。

【０２８８】

実施の形態４においては、話題抽出の側面を強調し、ある文書が複数の話題として分類される可能性を前提としており、たとえば、ネットワーク化に関する調査における「エンドユーザーがソフトウエアのインストール方法について聞いてくるのでシステム管理者としての仕事ができない」という回答について言えば、この意見は「ソフトウエアの操作方法理解に関する困難性」という話題として分類され得るし、「システム管理者の仕事の多忙さ」という話題で分類される可能性もある。

【０２８９】

実施の形態４においては、いずれにしても、「ソフトウエアの操作方法理解に関する困難性」というクラスタと「システム管理者の仕事の多忙さ」というクラスタの両方とも抽出したいというニーズに応えている。

【０２９０】

これとは反対に、操作者は、一度抽出した話題は既知であるので、次回の分類の際はなるべく異なる分類結果が欲しいとするケースも考えられる。実施の形態５では、このような要求に応えるため、ベクトル修正部３７０２により、ｎ回目の分類で選択されたクラスタに所属する文書のすべてまたはその一部を次回以降の分類を実行する際、分類対象から除去するものである。

【０２９１】

クラスタ選択指示部３０１０により選択指示を受けたクラスタの所属文書はベクトル記憶部３７０１において列ベクトルの形式で記憶されているため、ベクトル修正部３７０２では劣ベクトルを除去することで、次回以降の分類実行用の分類対象文書集合を生成する。

【０２９２】

さらに、実施の形態４と同様に、選択されたクラスタにより文書表現空間修正部３７０３は、ベクトル記憶部３７０１に記憶されている行列から特徴次元を削除する。

【０２９３】

つぎに、実施の形態５の文書分類装置の一連の処理の手順について説明する。図３８は、実施の形態５よる文書分類装置の一連の処理の手順を示すフローチャートである。図３８のフローチャートにおいて、まず、分類の対象となる文書が入力される（ステップＳ３８０１）。

【０２９４】

つぎに、入力された文書の言語が解析され（ステップＳ３８０２）、解析された結果、すなわち、抽出された単語に基づいて、文書特徴ベクトルが生成され（ステップＳ３８０３）、生成された文書特徴ベクトルが記憶される（ステップＳ３８０４）。

【０２９５】

その後、分類パラメータの指示があるのを待って、分類パラメータの指示があった場合（ステップＳ３８０５肯定）は、指示があった分類パラメータにしたがって文書を分類し（ステップＳ３８０６）、その結果、すなわち、クラスタに関する情報を記憶する（ステップＳ３８０７）。

【０２９６】

つぎに、分類されたクラスタの特徴を算出し（ステップＳ３８０８）、算出された結果を表示する（ステップＳ３８０９）。表示されたクラスタの中から、クラスタが選択されたか否かを判断し（ステップＳ３８１０）、選択されなかった場合（ステップＳ３８１０否定）は、ステップＳ３８０５へ移行し、再度分類パラメータの指示があるのを待つ（ステップＳ３８０５）。

【０２９７】

一方、ステップＳ３８１０において、クラスタが選択された場合（ステップＳ３８１０肯定）は、選択されたクラスタに関して分類体系を生成し、記憶する（ステップ３８１１）。この際、操作者により入力されたクラスタに関する情報を併せて記憶することもできる。その後、繰り返し処理をおこなう旨の指示があったか否かを判断する（ステップＳ３８１２）。

【０２９８】

ステップＳ３８１２において、繰り返して処理をおこなう旨の指示があった場

合（ステップＳ３８１２肯定）は、選択されたクラスタに所属する文書のすべてあるいはその一部の文書を除くように文書特徴ベクトルを修正する（ステップＳ３８１３）。

【０２９９】

ステップＳ３８１３に引き続き、「特徴次元（単語）」×「文書」の行列から判別された特徴次元について行ベクトルを削除することにより文書表現空間を修正する（ステップＳ３８１４）。その後、ステップＳ３８０５へ移行し、以後、ステップＳ３８０５～Ｓ３８１４を繰り返しおこなう。

【０３００】

一方、ステップＳ３８１２において、繰り返して処理をおこなう旨に指示がない場合（ステップＳ３８１２否定）は、これにより、一連の処理をすべて終了する。

【０３０１】

以上説明したように、実施の形態５よる文書分類装置によれば、ベクトル修正部３７０２が、既知になったクラスタの影響を排除し、かつ、文書表現空間修正部３７０３が、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができる。

【０３０２】

〔実施の形態６〕

さて、上述した実施の形態２または実施の形態４においては、繰り返し分類処理をおこなった場合に、ある文書が何度選択されたかその情報については考慮していなかったが以下に説明する実施の形態６のように、選択情報付与部を含む構成とし、選択情報をクラスタ特徴とともに表示するようにしてもよい。

【０３０３】

実施の形態６による文書分類装置を構成する情報処理システムは、図１に示したように実施の形態１と同様であるので、その説明は省略する。また、サーバー１０１およびクライアント１０２のハードウエア構成についても、図２・図３に示したように実施の形態１と同様であるので、その説明は省略する。

【０３０４】

つぎに、実施の形態６による文書分類装置の機能的構成について説明する。図３９は、この発明の実施の形態６による文書分類装置の構成を機能的に示すブロック図である。図３９において、実施の形態４の図３５と同一のものに関しては同じ符号を付して、その説明を省略する。

【０３０５】

図３９のブロック図において、文書分類装置は、入力部３００１、言語解析部３００２、ベクトル生成部３００３、分類部３００４、分類パラメータ指示部３００５、分類結果記憶部３００６、クラスタ特徴表示部３００７、クラスタ特徴算出部３００８、分類体系記憶部３００９、クラスタ選択指示部３０１０、分類体系閲覧操作部３０１１、ベクトル記憶部３５０１、文書表現空間修正部３５０２のほか、選択情報付与部３９０１を含む構成である。

【０３０６】

選択情報付与部３９０１は、分類部３００４により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与する。また、クラスタ特徴表示部３００７は、クラスタ特徴を表示するとともに、選択情報付与部３９０１により付与された選択情報を表示する。

【０３０７】

なお、選択情報付与部３９０１は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令にしたがってＣＰＵ２０１または３０１等が命令処理を実行することにより、機能を実現する。

【０３０８】

つぎに、選択情報付与部３９０１の詳細な内容について説明する。アンケートの調査の例において、独自性の高いユニークな意見は貴重であることが経験的に知られている。これは、調査を企画する担当者が予想できなかった意見である場合が多いからである。

【０３０９】

そこで、操作者に選択されたクラスタに所属する文書を、次回以降の分類実行の際に使用する場合において、クラスタ特徴表示部３００７で個々の文書を表示する際に、各文書が何回選択されたかを示すことで、多重に利用される文書の識別性を向上させ、かつ一度も選択されない文書の識別性も向上させることができる。

【０３１０】

図４０は、実施の形態６による文書分類装置の分類結果記憶部３００６において設けられたテーブル４０００を示す説明図である。図４０において、文書ＩＤごとにテーブル化されており、テーブル４０００は、各文書がどのサイクルに分類実行の際に操作者に選択されたかを記録する。すなわち、選択された場合は選択情報として「１」を記録し、選択されなかった場合は選択情報として「０」を記録する。

【０３１１】

たとえば、４回分類が実行された際、文書ＩＤの「１」、第１回目および第２回目の分類実行時に操作者に選択されたことを示し、第３回目、第４回目の分類実行時には選択されなかったことを示している。一方、文書ＩＤの「２」は、未だ一度も選択されておらず、操作者にとって未知の意見という可能性を示唆している。

【０３１２】

こうした情報に基づいて、クラスタ特徴表示部３００７が文書を操作者に表示する際、たとえば、選択された回数に応じて表示を変化させるようにするとよい。変化させる視覚的特性としては、たとえば文字や背景の色の濃度や彩度等が考えられる。

【０３１３】

また、直接的に数字やグラフ等で選択された回数を表現することもできる。いずれにしてもよ選択される文書と一度も選択されていない文書とを視覚的に識別できる表示形式であれば、上記のものに限らない。

【０３１４】

また、上記選択情報を分類体系閲覧操作部３０１１の閲覧操作により閲覧でき

るようにしてもよい。

【０３１５】

つぎに、選択情報付与部３９０１の処理の内容について説明する。図４１は、実施の形態６による文書分類装置の選択情報付与部３９０１の処理の手順を示すフローチャートである。図４１のフローチャートにおいて、まず、分類処理がおこなわれ（ステップＳ４１０１）、それに引き続き、最初の文書が抽出される（ステップＳ４１０２）。

【０３１６】

抽出された文書が、ステップＳ４１０１における分類処理の際に選択されたか否かを判断する（ステップＳ４１０３）。ここで、選択された場合（ステップＳ４１０３肯定）は、選択情報としてデータ「１」を記録する（ステップＳ４１０４）。一方、選択されなかった場合（ステップＳ４１０３否定）は、選択情報としてデータ「０」を記録する（ステップＳ４１０５）。

【０３１７】

つぎに、すべての文書について処理が終了したか否かを判断する（ステップＳ４１０６）。ここで、すべての文書について処理が終了していない場合（ステップＳ４１０６否定）は、つぎに文書を抽出し（ステップＳ４１０７）、ステップＳ４１０３へ移行し、以後、ステップＳ４１０３～Ｓ４１０７を繰り返しおこなう。

【０３１８】

一方、ステップＳ４１０６において、すべての文書について処理が終了した場合（ステップＳ４１０６肯定）は、ステップＳ４１０１へ移行し、再度分類処理がおこなわれる（ステップＳ４１０１）。このようにして、分類処理がおこなわれる回数だけ、ステップＳ４１０１～Ｓ４１０７の各処理が繰り返しおこなわれる。

【０３１９】

以上説明したように、実施の形態６によれば、選択情報付与部３９０１が選択情報を付与し、その選択情報をクラスタ特徴表示部３００７が表示するので、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させ

ることができる。

【０３２０】

なお、実施の形態２～５で説明した文書分類方法は、あらかじめ用意されたプログラムをパーソナルコンピュータやワークステーション等のコンピュータで実行することにより実現される。このプログラムは、ハードディスク、フロッピーディスク、ＣＤ－ＲＯＭ、ＭＯ、ＤＶＤ等のコンピュータで読み取り可能な記録媒体に記録され、コンピュータによって記録媒体から読み出されることによって実行される。またこのプログラムは、上記記録媒体を介して、または伝送媒体として、インターネット等のネットワークを介して配布することができる。

【０３２１】

つぎに、実施の形態７～１６に係る情報分類装置について説明する。本発明の実施の形態では、自然言語で記述された一つ以上の文の集まりであり、かつその一つ以上の文の集まりが分類される対象である場合、それを文書と言う。具体的な例をあげれば、ＩＰＣ分類等により分類される公開特許公報や、政治・経済・文化・科学技術等の特定分野に分類される新聞記事も文書であるし、それらから請求項や特定の一文を取り出したものであっても、請求項という分類に含まれる文であるか、用途等により分類可能な特定の一文であれば文書とみなす。以下、図面によりこの発明の実施の形態７～１６を詳細に説明する。

【０３２２】

〔実施の形態７〕

図４２はこの発明の実施の形態７を示す文書分類装置の構成ブロック図である。図４２に示したように、実施の形態７の文書分類装置は、文書データ群を入力する文書入力部（文書入力手段）５００１、それぞれの文書データを所定の基準に基づいて一つまたは複数の分割文書データに分割する文書分割部（文書分割手段）５００２、上記文書データと分割文書データとを対応付けるマップを生成する文書－分割文書対応マップ生成部（文書－分割文書対応マップ生成手段）５００３を備えている。

【０３２３】

また、上記文書分類装置は、分割文書データつまり分割された文書を分類する

分割文書分類部（分割文書分類手段）５００４、分割文書分類結果情報を生成する分割文書分類結果生成部（分割文書分類結果生成手段）５００５、上記文書－分割文書対応マップと上記分割文書分類結果情報とを用いて上記文書データの分類結果情報を生成する文書分類結果生成部（文書分類結果生成手段）５００６などを備えている。

【０３２４】

なお、上記文書分割部５００２、文書－分割文書対応マップ生成部５００３、分割文書分類部５００４、分割文書分類結果生成部５００５、文書分類結果生成部５００６は共有または独自のプログラム記憶用メモリおよびプログラムにしたがって動作するＣＰＵを有している。

【０３２５】

以下、図４２などにしたがって、実施の形態７の文書分類装置、文書分類方法を詳細に説明する。まず、文書入力部５００１により、文書群が入力される。上記文書入力部５００１はキーボード、ＯＣＲ装置、着脱型記録媒体、またはネットワーク通信手段を備え、それらのいずれか一つを介して文書データ群を入力するのである。

【０３２６】

そして、文書分割部５００２が上記文書データ群を取得し、それぞれの文書データを所定の基準に基づいて分割し、一つの文書データから一つまたは複数の分割文書データを生成する。なお、文書データを分割する方法としては、文書の構造情報や文書を構成する要素情報を用いたり、利用者が指定する方法などを用いるが、ここでは、その方法は問わないこととする。

【０３２７】

図４３に、この文書分類装置／文書分類方法でおこなわれる、文書データから複数の分割文書データを生成する一例を示す。この例に示した文書１には複数のニューストピックが記述されており、１日分のトピックが文書単位となっている。図示したように、この文書ではそれぞれのニューストピックが二つの改行コードにより分離されているので、この規則を用いて一つの文書である文書１を分割し、一つが一つのトピックにより形成される分割文書１－１～１－７の７つの分

割文書データを生成する。なお、分割前の文書１も分割文書データとして含めることもできるが、ここでは含めないことにする。

【０３２８】

文書が分割されると、文書－分割文書対応マップ生成部５００３が分割前の文書データとその文書データから生成された分割文書データとを対応付けるマップを生成する。たとえば、個々の文書データを一意に示す識別子と個々の分割文書データを一意に示す識別子とから構成されるマップ、あるいは文書データごとに分割文書データを一意に示す識別子からなるマップを生成するのである。なお、文書データと分割文書データを対応付ける方法についてはここでは問わないこととする。

【０３２９】

図４４に、文書－分割文書対応マップを生成する一例を示す。図４４において、文書１～文書３は文書データを示し、分割文書１～分割文書１２は分割文書データを示している。図示のように、それぞれの文書データおよび分割文書データにそれぞれを一意に識別することかできる識別番号（識別子）を付与し、上記文書データの識別番号と分割文書データの識別番号とを図４４の左下に示したテーブル形式で対応づけている。なお、任意の複数の分割文書データが文書分類にて用いられる基準において同一とみなすことができる場合は、それらの識別番号を同一にしてもよい。

【０３３０】

続いて、分割文書分類部５００４が上記分割文書を対象に文書分類をおこなう。個々の分割文書に対して、たとえば、言語処理を施し、文書中に含まれているそれぞれの単語の出現頻度を計数し、それに基づいてそれぞれの文書の特徴を計量的に表す特徴ベクトルを求め、それらの特徴ベクトルに対してカイ自乗法、判別分析手法、またはクラスタ分析手法などを適用することにより文書分類をおこなう。

【０３３１】

つぎに、図４５に示すように、分割文書分類結果生成部５００５が上記の分割文書分類の結果に基づいた分割文書分類結果情報を生成する。

【０３３２】

ここで、分割文書分類結果情報とは、たとえば、各分割文書データの所属カテゴリに関する情報（たとえば、図４５に示した「分割文書データを３つのカテゴリに分類した結果」という表中の「分類カテゴリ」および「所属カテゴリの代表値との距離」の項の情報）、生成された所属カテゴリ個々に関する情報（たとえば、図４５に示した「分類カテゴリに関する情報」という表中の「代表値」および「所属データ数（分割文書数）」の項の情報）、生成された所属カテゴリ間の情報（たとえば図４５に示した「分類カテゴリ間の距離」という表の中の情報）などである。なお、利用者は上記のような種々の情報を分類結果分析の際の基礎データとして利用することができる。

【０３３３】

図４５は、１２個の分割文書データをそれらの有する計量的特徴ベクトルを用いて３つのカテゴリに分類した場合の分類結果の生成例である。分割文書データの有する計量的な３次元ベクトル（ベクトルの成分数は分類対象文書群に生起するすべての単語の種類数になるが、ここでは、いくつかの単語が縮退した３次元ベクトルに線形変換している）に対してたとえばクラスタ分析手法の一つであるＷａｒｄ法などを適用することで３つのカテゴリに分類することができる。

【０３３４】

つまり、各分割文書データは図示したように３つのカテゴリのうちのいずれか一つに属する。なお、所属カテゴリの代表値とは、所属分割文書データの特徴ベクトルの平均値（所属分割文書データの重心）である。

【０３３５】

また、所属カテゴリの代表値との距離（類似度に対応する）は、たとえば、図４５の分割文書３については、分割文書データ特徴ベクトルの項における分割文書３の値と、分割文書３の分類カテゴリであるカテゴリ２の代表値（所属分割文書データの重心）の項の値により、以下の数式から求めることができる。

【０３３６】

$$((3.00-2.66)^2+(2.00-2.00)^2\div(4.00-3.66)^2)^{1/2}=0.48$$

上記の所属カテゴリの代表値との距離が小さいほど、そのカテゴリに属する平均的分割文書との類似度が高いということになる。

【０３３７】

なお、分割文書分類結果情報としては、図４５に示した以外にも、カテゴリ内分散やカテゴリ間分散、各カテゴリにおける類似度のレンジなどさまざまな統計量を生成することかできる。

【０３３８】

続いて、文書分類結果生成部５００６が上記文書－分割文書対応マップと上記分割文書分類結果情報とを用いて、たとえば図４６に示すような、上記文書データの分類結果情報を生成する。図４６の例では、図示したように、各分類カテゴリごとに、所属する分割文書データ、その類似度（所属カテゴリの代表値との距離）、分割文書データの属する分割前文書データ（所属文書）、文書占有率（分割文書データの当該カテゴリに所属する割合）、分割文書データの所属文書における相対位置（順序）、所属カテゴリ内での当該分割文書データの類似度の順位などを生成している。

【０３３９】

なお、上記において、所属文書は文書－分割文書対応マップから、それ以外の分類結果情報は分割文書分類結果情報から得ている。文書分類結果生成部５００６は図４６に示した情報以外にも、各カテゴリ内での分散、分割文書データの所属カテゴリ内での偏差値などさまざまな統計量、文書データや分割文書データの内容などを分類結果情報として利用することもできる。

【０３４０】

また、上記においては、すべての結果を分割文書データを単位とした表形式で表現しているが、分類カテゴリや文書データを単位として表現することもできる。また、分類結果情報をテキスト表現にするだけでなく、グラフィカルな表現にして、利用者が理解しやすいようにすることも可能である。

【０３４１】

こうして、本実施の形態によれば、一つの文書が分割され、分割文書が分類され、分割前文書と上記分割文書との対応が利用者に示され、上記分割文書の分類

結果が利用者に示されるので、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりすることがなく、したがって、利用者がその分類カテゴリをよく理解できる。また、分割前文書（所属文書）中の分割文書の位置なども示されるので、利用者は文書群中の読みたい部分を効率的に読むことができる。

【０３４２】

〔実施の形態８〕

図４７は本発明の実施の形態８に係る文書分類装置の構成ブロック図である。図示したように、実施の形態８の文書分類装置は、図４２に示した実施の形態７の構成に加え、文書データを保存する文書保存部（文書保存手段）５００７、分割文書データを保存する分割文書保存部（分割文書保存手段）５００８、文書－分割文書対応マップ生成部５００３により生成された文書－分割文書対応マップを保存する文書－分割文書対応マップ保存部（文書－分割文書対応マップ保存手段）５００９を備えている。なお、上記各保存部はたとえば共有のハードディスクや半導体メモリなどにより構成される。

【０３４３】

上記した構成により、本実施の形態の文書保存部５００７は、文書データの内容や、文書の作成者、作成日、最終修正日などの文書データに付随する情報を適切な形式で保存する。また、文書データが文書内容とともにその要素からなる計量的な特徴ベクトルを持つ場合にはこれらも保存する。文書入力部５００１にて、個々の文書データにそれらを一意に表す識別子が付与される場合にはこの識別子も適切な形式で保存することができる。

【０３４４】

また、分割文書保存部５００８は、文書分割部５００２により生成される分割文書データの内容を適切な形式で保存するとともに、計量的な特徴ベクトルを持つ場合にはこれらも保存する。個々の上記分割文書データにそれらを一意に表す識別子が付与される場合にはこの識別子も適切な形式で保存することができる。

【０３４５】

また、文書－分割文書対応マップ保存部５００９は、文書－分割文書対応マップ生成部５００３により生成される文書－分割文書対応マップを適切な形式で保存する。

【０３４６】

このように、実施の形態８によれば、文書データ、分割文書データ、および文書－分割文書対応マップが保存されるので、分割文書データおよび文書－分割文書対応マップを再生成することなしに、同一の文書データに対して、分類数、分類手法、または分類時の諸設定などパラメータの異なる分類結果を効率的に求めることができる。また、文書データを分類し、分類結果を生成するために必要なデータが保存されることにより、利用者は、分類作業に対して時間的な自由度を持つことができ、過去に行った文書分類の再分析を任意の時間におこなうこともできる。

【０３４７】

〔実施の形態９〕

図４８は本発明の実施の形態９を示す文書分類装置の構成ブロック図である。図４８に示したように、本実施の形態の文書分類装置は、図４７に示した実施の形態８の構成に加え、分割文書分類結果生成部５００５により生成された分割文書分類結果情報を保有する分割文書分類結果保存部（分割文書分類結果保存手段）５０１０を備えている。なお、上記分割文書分類結果保存部５０１０は、たとえば、共有のハードディスクや半導体メモリなどにより構成される。

【０３４８】

このように、第３の実施の形態によれば、文書データ、分割文書データ、文書－分割文書対応マップ、および、分割文書分類結果情報が保存されるので、実施の形態８の効果に加え、一度分類を実行すれば、その分類結果をテキスト表現や表表現やグラフ表現などさまざまな形式で表現することかできる。また、分割文書分類結果情報が保存されることにより、分類の実行作業および分類結果の分析作業において、利用者は、時間的な自由度を持つことができ、過去に行った文書分類結果の再分析をさまざまな表現形式で任意の時間におこなうこともできる。

【０３４９】

〔実施の形態１０〕

この発明の実施の形態１０では、前記各実施の形態の文書分類装置、文書分類方法において、図４９に示すように、文書分割部５００２により生成される複数の分割文書データ中に分割前の文書データである文書１を含む。これにより、本実施の形態では、利用者は、分割されている文書データを分類することで得られる詳細な文書データの分類構造だけでなく、分割前の文書データ自体を分類した結果として得られるマクロな分類構造の融合した分類構造を得ることができる。

【０３５０】

〔実施の形態１１〕

この発明の実施の形態１１では、前記各実施の形態の文書分類装置、文書分類方法において、文書分割部５００２は、文書データの構造情報を基に文書データを分割する。図５０に、分類対象文書データかＨＴＭＬ形式で記述された文書の例を示す。分割をおこなう前に、図５０に示したようなＨＴＭＬ形式の文書データから構造情報を抽出し、それらの構造を用いて文書の適切な分割規則を設定することにより文書データから分割文書データを生成する。

【０３５１】

つまり、この例では、文書データ中のタグ＜Ｌ１＞に着目し、「タグ＜Ｌ１＞を持つテキストを一つの分割文書データとする」という文言を分割文書データを生成する規則とする。この規則を文書データに適用することにより図５０に示したような７つの分割文書か生成される。

【０３５２】

上記のように、文書が、ＨＴＭＬ、ＸＭＬ、ＳＧＭＬなど特定の構造化文書の形式を有していない場合でも、文字の大きさ、文字の装飾、文字の色、およびフォントなどに関する情報から分割規則を生成し、分割文書を生成することもできる。また、文書データがイメージであってＯＣＲ装置などにより入力される場合には、元のイメージのレイアウト情報などを利用することにより分割規則を生成し、分割文書を生成することもできる。

【０３５３】

なお、文書データのすべてをいずれかの分割文書データにする必要はない。た

とえば、図５０に示した例では、文字列「ニューストピック（９８／０９／２５）」は分割文書には採用しない。

【０３５４】

このように、実施の形態１１では、文書データから構造情報を抽出し、文書割をおこなう前に構造情報を用いて文書の適切な分割規則を設定することにより、異なった話題の分割などを適切におこなうことができ、したがって、文書データの詳細な分類構造がわかる文書分類を適切におこなうことができる。

【０３５５】

〔実施の形態１２〕

この発明の実施の形態１２では、前記実施の形態７～１０の文書分類装置、文書分類方法において、図５１に示すように、文書データに含まれる単語など要素を抽出する文書要素解析部（文書要素抽出手段）５０１１、上記文書要素解析部５０１１により抽出された要素に付随する品詞など要素付随情報を抽出する要素付随情報抽出部（要素付随情報抽出手段）５０１２を備え（図５１は図４８に示した実施の形態９に文書要素抽出部５０１１、要素付随情報抽出５０１２を加えた例で示している）、文書分割部５００２が、上記文書要素解析部５０１１により抽出された要素、または上記要素と上記要素付随情報抽出部５０１２により抽出された要素付随情報とを用いて上記文書データを分割する。

【０３５６】

図５２に示すように、文書分割をおこなう前に、自然言語処理手段である文書要素解析部５０１１が文書データから単語などそれらの要素を抽出し、要素付随情報抽出部５０１２が品詞など要素付随情報を抽出して文書の適切な分割規則を設定するのである。なお、上記文書要素解析部５０１１および要素付随情報抽出部５０１２は新たに設けるのではなく、分割文書分類部５００４内の同様の手段を用いることが可能である。

【０３５７】

この実施の形態では、たとえば、図５２に示したように、文書データが特定の構造情報を持たない複数のニューストピックの集まりであり、各トピックが、単語「トピック」＋「数字」＋「改行コード」という文字列の後に記述されている

場合で説明すると、上記のような構造が文書要素解析部５０１１および要素付随情報抽出部５０１２の抽出結果から認識され、文章の終端を考慮して、「トピック＋数字＋改行コードという文字列を先頭とし、上記文字列または文書終端記号を終端として囲まれる文字列を一つの分割文書データとする」という分割規則が生成されることになる。

【０３５８】

さらに詳しく説明すると、抽出された単語とその品詞情報などから、まず、名詞と改行コードのみを抽出し、つぎに、文字列「トピック＋数字＋改行コード」および文書終端記号を検出し、文書内でのそれらの位置を記憶する。そして、文書データに対して前記分割規則を適用し、図５２に示したような分割文書データを生成する。

【０３５９】

なお、文書データのすべてをいずれかの分割文書データにする必要はなく、たとえば、図５２に示した例では、文字列「ニューストピック（９８／０９／２５）」は分割文書には採用しない。また、上記の例では、文書データから要素およびその付随情報を抽出して分割規則を設定する場合で説明したが、要素のみを抽出してその要素情報から分割規則を設定することも可能である。

【０３６０】

こうして、実施の形態１２によれば、文書データからそれらの要素情報などを抽出し、抽出した要素情報などを用いて文書の分割規則を設定することにより、実施の形態１１と同様に、文書データの詳細な分類構造がわかる文書分類を適切におこなうことができる。

【０３６１】

〔実施の形態１３〕

この発明の実施の形態１３では、前記実施の形態７～１０の文書分類装置、文書分類方法において、利用者により指示された指定範囲にしたがって文書分割部５００２が文書データを分割する。図５３に示すような文書データに対して利用者がそれぞれの分割文書の範囲を指定すると、指定にしたがって文書分割部５００２が文書分割をおこなう。

【０３６２】

本実施の形態では、文書分割時、文書分割部５００２がまず、画面上に、その初期状態として左右の指示ポイントおよび領域指定ラインからなる領域指定オブジェクトを文書の最上部に表示する。この状態で、利用者は、マウスなどポインティングデバイスを用いて、左右どちらかの指示ポイントをドラッグし、それを上下に移動させることにより、それぞれの分割文書の領域を選択することができる。

【０３６３】

また、この指定時、文書分割部５００２は、領域選択処理をおこなっていることを示すため、指示ポインタを黒色から白色に、領域指定ラインを実線から破線に変化させる。選択領域を決定するには、所望の位置で指示ポイントのドラッグを止めればよい。

【０３６４】

つぎに、利用者は選択した領域を分割文書とするかしないか決定する。分割領域としない場合には、それを明示的に表示するために、文書分割部５００２は選択領域を図示のように網掛け表示にさせる。

【０３６５】

こうして、本実施の形態によれば、利用者は文書データからそれぞれの分割文書データを所望通りに選択することができるので、文書データの詳細な分類構造がわかり、かつ利用者の意図に合った文書分類をおこなうことができる。

【０３６６】

〔実施の形態１４〕

この発明の実施の形態１４では、前記実施の形態７～１０の文書分類装置、文書分類方法において、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する。たとえば、図５４に示す文書データをほぼ２００文字を単位として分割をおこなう。

【０３６７】

ここで、ほぼ２００文字を単位とするのは、正確な２００文字単位としてもその終端が句点である保証がないことから、２００文字目の前後のもっとも近い句

点をそれぞれの分割文書の終端とするからである。こうして、図５４に示したような分割文書が生成される。同様に、所定の文数を単位とした文書分割をおこなうこともできるし、文字数と文数の両方を基にした文書分割をおこなうこともできる。

【０３６８】

このように、実施の形態１４によれば、文字数、文数、または文字数と文数の両方を基に文書データを分割することにより、話題の異なった内容などが異なった分割文書として分割され、分類される可能性が高くなるので、文書データの詳細な分類構造がわかる文書分類をおこなうことができる。

【０３６９】

〔実施の形態１５〕

この発明の実施の形態１５では、前記各実施の形態の文書分類装置、文書分類方法において、文書分類結果生成部５００６が分類結果情報として、文書データを示す情報および上記文書データに付随する代表的情報のみを提示する。

【０３７０】

たとえば図５５に示すように、先頭に分類カテゴリ名を表示し、その横にそのカテゴリを代表するキーワードを表示し、カテゴリ名の下には文書データを示す情報として当該カテゴリに属する分割文書データを含んでいる文書データの、たとえば、文書データ名（文書名）を表示する。また、各文書データ名の左側には文書アイコンを表示させ、この文書アイコンが指示されたとき、文書データの内容を表示させる。

【０３７１】

また、各文書データ名の配置順は、カテゴリ代表値との類似度が高い分割文書データの文書データ名を先（左側）にする。また、同じ文書データから生成された複数の分割文書データが同一の分類カテゴリに属している場合には、類似度のもっとも高い分割文書データに対応する文書データ名のみを表示する。なお、上記キーワードとは出現頻度の多い単語である。

【０３７２】

このように、実施の形態１５によれば、文書分類結果が文書データを示す情報

と文書データに付随する代表的情報のみが表示されるので、利用者は文書データの詳細な分類構造の概要を容易に把握することができる。

【０３７３】

〔実施の形態１６〕

この発明の実施の形態１６では、実施の形態１５の文書分類結果提示に加えて、分割文書データを示す情報および上記分割文書データに付随する情報を提示する。

【０３７４】

たとえば、図５６に示すように、先頭に分類カテゴリ名を表示し、その横にそのカテゴリを代表するキーワードを表示し、カテゴリ名の下には文書データを示す情報として当該カテゴリに属する分割文書データを含んでいる文書データのたとえば文書データ名（文書名）を表示する。

【０３７５】

また、各文書データ名の左側には文書アイコンを表示させ、この文書アイコンが指示されたとき、文書データの内容を表示させる。また、文書データ名の右側には分割文書アイコンを表示させる。なお、分割文書アイコン中には当該文書データにおける分割文書データの位置と当該文書データ中の分割文書数を表示させる。さらに、上記分割文書アイコンを指示することで文書データ中の当該分割文書データを表示させることができる。

【０３７６】

また、各文書データ名の配置順はカテゴリ代表値との類似度が高い分割文書データの文書データ名を先にする。また、同じ文書データから生成された複数の分割文書データが同一の分類カテゴリに属している場合には類似度の順位がわかるようにその順位を表示させる。

【０３７７】

このように、実施の形態１６によれば、文書分類結果が文書データを示す情報と文書データに付随する代表的情報、および分割文書データを示す情報と分割文書データに付随する代表的情報のみが表示されるので、利用者は文書データの詳細な分類構造の概要とともにどの分割文書が起因して当該カテゴリに分類された

かというようなことも容易にわかる。

【０３７８】

以上、本発明の文書分類装置および文書分類方法を説明したが、この文書分類方法を実現するプログラムを着脱可能であるとともにコンピュータ読み取り可能な記録媒体に記録し、上記記録媒体を移した先の情報処理装置内で本発明によった文書分類をおこなうこともできる。

【０３７９】

【発明の効果】

以上説明したように、請求項１の発明によれば、入力された文書データを記憶する文書記憶手段と、前記文書記憶手段により記憶された文書データの全部または一部を選択する選択手段と、前記選択手段により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出手段と、前記特徴抽出手段により抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理手段と、前記加工処理手段により加工処理された文書データの全部または一部を出力する出力手段とを備えるため、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【０３８０】

また、請求項２の発明によれば、前記出力手段が、前記加工処理手段により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定手段と、前記項目値設定手段により設定された項目値ごとに前記文書データの全部または一部を集計する集計手段と、を備え、前記文書データの全部または一部を、項目値を少なくとも一つの軸とする表形式に展開して出力するため、簡易な操作で加工処理の結果をクロス表として表すことができ、情報の内容の把握を容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【０３８１】

また、請求項３の発明によれば、前記出力手段が、さらに、前記加工処理手段により加工処理された文書データの全部または一部を、前記加工処理手段により加工処理される前の文書データの全部または一部とともに出力するため、加工処理すべき対象データとその他のデータが同時に表示され、それを確認することにより、加工処理の対象範囲の決定を正確かつ容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【０３８２】

また、請求項４の発明によれば、前記文書記憶手段が、さらに、前記加工処理手段により加工処理された文書データの全部または一部を記憶するため、以後、他のデータと同様に扱うことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【０３８３】

また、請求項５の発明によれば、前記選択手段が、さらに、前記出力手段により出力された文書データの全部または一部を選択するため、出力手段により出力された文書データの全部または一部をさらなる分析の対象とすることができ、多彩で高度な情報分析作業ができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという効果を奏する。

【０３８４】

また、請求項６の発明によれば、前記文書記憶手段が、さらに、前記加工処理の内容に関するデータを記憶するため、加工処理の内容に関するデータの紛失を防止し、当該データの管理が容易になるだけでなく、加工処理に用いた設定とそれによる処理結果を関連づけて把握することができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理装置が得られるという

効果を奏する。

【０３８５】

また、請求項７の発明によれば、入力手段が、文書データを入力し、言語解析手段が、前記入力手段により入力された文書データを解析して言語解析情報を得、ベクトル生成手段が、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類手段が、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出手段が、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、分類体系記憶手段が、前記クラスタ特徴算出手段により算出されたクラスタ特徴を分類体系の構成要素として記憶するため、クラスタを得ることができるとともに、クラスタ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０３８６】

また、請求項８の発明によれば、入力手段が、文書データを入力し、言語解析手段が、前記入力手段により入力された文書データを解析して言語解析情報を得、ベクトル生成手段が、前記言語解析手段により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類手段が、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出手段が、前記分類手段により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、表示手段が、前記クラスタ特徴算出手段により算出されたクラスタ特徴を表示し、クラスタ選択指示手段が、前記分類手段により生成された文書の部分集合の中から所望の部分集合を選択し、分類体系記憶手段が、前記クラスタ選択指示手段により選択された文書の部分集合を分類体系の構成要素として記憶するため、選択されたクラスタのみを用いて、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０３８７】

また、請求項９の発明によれば、請求項８の発明において、文書特徴ベクトル記憶手段が、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶し、ベクトル修正手段が、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトルを、前記クラスタ選択指示手段により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正し、前記分類手段が、前記ベクトル修正手段により修正された文書特徴ベクトルに基づいて文書を分類するため、既知になったクラスタの影響を排除した新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０３８８】

また、請求項１０の発明によれば、請求項８の発明において、文書特徴ベクトル記憶手段が、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶し、文書表現空間修正手段が、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択された部分集合から算出する特徴量に基づいて修正し、前記分類手段が、前記文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル生成手段により生成された文書特徴ベクトル間の類似度に基づいて文書を分類するため、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０３８９】

また、請求項１１の発明によれば、請求項９の発明において、文書特徴ベクトル記憶手段が、前記ベクトル生成手段により生成された文書特徴ベクトルを記憶し、文書表現空間修正手段が、前記文書特徴ベクトル記憶手段により記憶された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示手段により選択されたクラスタ特徴に基づいて修正し、前記分類手段が、前記

文書表現空間修正手段により修正された文書表現空間を用いて、前記ベクトル修正手段により修正された文書特徴ベクトル間の類似度に基づいて文書を分類するため、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０３９０】

また、請求項１２の発明によれば、請求項８または１０の発明において、選択情報付与手段が、前記分類手段により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与し、前記表示手段が、前記クラスタ特徴を表示するとともに、選択情報付与手段により付与された選択情報を表示するため、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０３９１】

また、請求項１３の発明によれば、請求項８～１２の発明において、前記分類体系記憶手段が、前記選択指示手段により選択された文書の部分集合に属する全部あるいは一部の文書のほか、クラスタ特徴および／または操作者が作成した任意の情報を分類体系の構成要素として記憶するため、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できるので、分類体系の利用価値を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類装置が得られるという効果を奏する。

【０３９２】

また、請求項１４の発明によれば、文書の内容にしたがって文書群を分類する文書分類装置において、文書データ群を入力する文書入力手段と、入力された文書データ群の各文書に対して所定の基準に基づき文書の分割をおこない、一つの

文書データから一つまたは複数の分割文書データを生成する文書分割手段と、前記文書データと前記分割文書データとの対応を示す文書－分割文書対応マップを生成する文書－分割文書対応マップ生成手段と、前記分割文書データを分類する分割文書分類手段と、前記分割文書分類手段による分類結果に基づいて分割文書分類結果情報を生成する分割文書分類結果生成手段と、前記文書－分割文書対応マップと前記分割文書分類結果情報とを用いて前記文書データの分類結果情報を生成する文書分類結果生成手段と、を備えるため、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりすることがなく、したがって、利用者がその分類カテゴリをよく理解が可能で、また、分割前文書（所属文書）中の分割文書の位置なども示されるので、利用者は文書群中の読みたい部分を効率的に読むことが可能な文書分類装置が得られるという効果を奏する。

【０３９３】

また、請求項１５の発明によれば、請求項１４の発明において、前記文書データを保存する文書保存手段と、前記分割文書データを保存する分割文書保存手段と、前記文書－分割文書対応マップ生成手段により生成された文書－分割文書対応マップを保存する文書－分割文書対応マップ保存手段と、を備えるため、分割文書データおよび文書－分割文書対応マップを再生成することなしに、同一の文書データに対して、分類数、分類手法、または分類時の諸設定などパラメータの異なる分類結果を効率的に求めることが可能で、また、文書データを分類し、分類結果を生成するために必要なデータが保存されることにより、利用者が分類作業に対して時間的な自由度を持つことが可能で、過去に行った文書分類の再分析を任意の時間間におこなうことも可能な文書分類装置が得られるという効果を奏する。

【０３９４】

また、請求項１６の発明によれば、請求項１５の発明において、前記分割文書分類結果生成手段により生成された分割文書分類結果情報を保存する分割文書分類結果保存手段を備えるため、請求項１５の発明の効果に加え、一度分類を実行

すれば、その分類結果をテキスト表現や表表現やグラフ表現などさまざまな形式で表現することが可能で、また、分割文書分類結果情報が保存されることにより、分類の実行作業および分類結果の分析作業において、利用者が時間的な自由度を持つことが可能で、過去に行った文書分類結果の再分析をさまざまな表現形式で任意の時間におこなうことも可能な文書分類装置が得られるという効果を奏する。

【０３９５】

また、請求項１７の発明によれば、請求項１４～１６の発明において、前記文書分割手段により生成される複数の分割文書データには分割前の文書データそのものを含むため、利用者は、分割されている文書データを分類することで得られる詳細な文書データの分類構造だけでなく、分割前の文書データ自体を分類した結果として得られる概略的でマクロな分類構造の融合した分類構造を得ることが可能な文書分類装置が得られるという効果を奏する。

【０３９６】

また、請求項１８の発明によれば、請求項１４～１７の発明において、前記文書分割手段が、文書データの構造情報を基に文書データを分割する構成にしたため、異なった話題の分割等を適切におこなうことができ、したがって、文書データの詳細な分類構造がわかる文書分類を適切におこなうことが可能な文書分類装置が得られるという効果を奏する。

【０３９７】

また、請求項１９の発明によれば、請求項１４～１７の発明において、前記文書データに含まれる要素を抽出する文書要素抽出手段と、前記文書要素抽出手段により抽出された要素に付随する要素付随情報を抽出する要素付随情報抽出手段と、を備え、前記文書分割手段が、前記文書要素抽出手段により抽出された要素、または前記要素と前記要素付随情報抽出手段により抽出された要素付随情報とを用いて前記文書データを分割する構成にしたため、文書データの詳細な分類構造がわかる文書分類を適切におこなうことが可能な文書分類装置が得られるという効果を奏する。

【０３９８】

また、請求項２０の発明によれば、請求項１４～１７の発明において、前記文書分割手段が、指示された指定範囲にしたがって文書データの分割をおこなう構成にしたため、利用者の意図に合い、かつ文書データの詳細な分類構造がわかる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【０３９９】

また、請求項２１の発明によれば、請求項１４～１７において、前記文書分割手段が、文書データ中の文字数、文数、または文字数と文数の両方を基に文書データを分割する構成にしたため、話題の異なった内容などが異なった文書として分類される可能性が高くなり、したがって、この発明でも文書データの詳細な分類構造がわかる文書分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【０４００】

また、請求項２２の発明によれば、請求項１４～２１の発明において、前記文書分類結果生成手段が、文書データを示す情報および前記文書データに付随する代表的情報を、分類結果情報として抽出して提示する構成にしたため、利用者は文書データの詳細な分類構造の概要や全体的な構造を容易に把握することが可能な文書分類装置が得られるという効果を奏する。

【０４０１】

また、請求項２３の発明によれば、請求項２２の発明において、前記文書分類結果生成手段が、分割文書データを示す情報および前記分割文書データに付随する代表的情報を、分類結果情報として、抽出して提示する構成にしたため、利用者は文書データの詳細な分類構造の概要や全体的な構造とともにどの分割文書が起因して当該カテゴリに分類されたかというようなことも容易にわかる文書分類装置が得られるという効果を奏する。

【０４０２】

また、請求項２４の発明によれば、入力された文書データを記憶する文書記憶工程と、前記文書記憶工程により記憶された文書データの全部または一部を選択する選択工程と、前記選択工程により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出工程と、前記特徴抽出工程に

より抽出された文字列の特徴に関するデータに基づいて前記文書データの全部または一部を加工処理する加工処理工程と、前記加工処理工程により加工処理された文書データの全部または一部を出力する出力工程と、を含むので、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【０４０３】

また、請求項２５の発明によれば、前記出力工程が、前記加工処理工程により加工処理された文書データの全部または一部の内容に基づいて複数の項目値を設定する項目値設定工程と、前記項目値設定工程により設定された項目値ごとに前記文書データの全部または一部を集計する集計工程と、を含み、前記文書データの全部または一部を、項目値を少なくとも一つの軸とする表形式に展開して出力するので、簡易な操作で加工処理の結果をクロス表として表すことができ、情報の内容の把握を容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【０４０４】

また、請求項２６の発明によれば、前記出力工程が、さらに、前記加工処理工程により加工処理された文書データの全部または一部を、前記加工処理工程により加工処理される前の文書データの全部または一部とともに出力するので、加工処理すべき対象データとその他のデータが同時に表示され、それを確認することにより、加工処理の対象範囲の決定を正確かつ容易におこなうことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【０４０５】

また、請求項２７の発明によれば、前記文書記憶工程が、さらに、前記加工処理工程により加工処理された文書データの全部または一部を記憶するので、以後

、他のデータと同様に扱うことができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【０４０６】

また、請求項２８の発明によれば、前記選択工程が、さらに、前記出力工程により出力された文書データの全部または一部を選択するので、出力工程により出力された文書データの全部または一部をさらなる分析の対象とすることができ、多彩で高度な情報分析作業ができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【０４０７】

また、請求項２９の発明によれば、前記文書記憶工程が、さらに、前記加工処理の内容に関するデータを記憶するので、加工処理の内容に関するデータの紛失を防止し、当該データの管理が容易になるだけでなく、加工処理に用いた設定とそれによる処理結果を関連づけて把握することができることから、文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことが可能な文書処理方法が得られるという効果を奏する。

【０４０８】

また、請求項３０の発明によれば、入力工程が、文書データを入力し、言語解析工程が、前記入力工程により入力された文書データを解析して言語解析情報を得、ベクトル生成工程が、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出工程が、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、分類体系生成工程が、前記クラスタ特徴算出工程により算出されたクラスタ特徴に基づいて分類体系の構成要素を生成するので、クラスタを得ることができるとともに、クラス

タ重心間の類似度等を用いて、クラスタの内容に基づくクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０４０９】

また、請求項３１の発明によれば、入力工程が、文書データを入力し、言語解析工程が、前記入力工程により入力された文書データを解析して言語解析情報を得、ベクトル生成工程が、前記言語解析工程により得られた言語解析情報に基づいて前記文書データに対する文書特徴ベクトルを生成し、分類工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類し、文書の部分集合を生成し、クラスタ特徴算出工程が、前記分類工程により生成された文書の部分集合の特徴であるクラスタ特徴を算出し、表示工程が、前記クラスタ特徴算出工程により算出されたクラスタ特徴を表示し、クラスタ選択指示工程が、前記分類工程により生成された文書の部分集合の中から所望の部分集合を選択し、分類体系生成工程が、前記クラスタ選択指示工程により選択されたクラスタ特徴に基づいて分類体系の構成要素を生成するので、選択されたクラスタのみを用いて、より操作者の意図したものに近いクラスタの構造化・体系化をおこなうことができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０４１０】

また、請求項３２の発明によれば、請求項３１の発明において、ベクトル修正工程が、前記クラスタ選択指示手段により選択された部分集合に属する文書の文書特徴ベクトルを除去したのこりとなるように修正し、前記分類工程が、前記ベクトル修正工程により修正された文書特徴ベクトルに基づいて文書を分類するので、既知になったクラスタの影響を排除した新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０４１１】

また、請求項３３の発明によれば、請求項３１の発明において、文書表現空間修正工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特徴量に基づいて修正し、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル生成手段工程により生成された文書特徴ベクトル間の類似度に基づいて文書を分類するので、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０４１２】

また、請求項３４の発明によれば、請求項３２の発明において、文書表現空間修正工程が、前記ベクトル生成工程により生成された文書特徴ベクトル間の類似度を判断する際の文書表現空間を前記クラスタ選択指示工程により選択された部分集合から算出する特徴量に基づいて修正し、前記分類工程が、前記文書表現空間修正工程により修正された文書表現空間を用いて、前記ベクトル修正工程により修正された文書特徴ベクトル間の類似度に基づいて文書を分類するので、既知になったクラスタの影響を排除し、かつ、前回の分類実行の結果、操作者に選択されたクラスタの形成特徴を次回の分類実行時に排除することができ、排除した状態で新たなクラスタを生成することができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０４１３】

また、請求項３５の発明によれば、請求項３１または３３の発明において、選択情報付与工程が、前記分類工程により生成された文書の部分集合に所属する文書のすべてあるいは一部が選択された場合に選択されたことを示す選択情報を付与し、前記表示工程が、前記クラスタ特徴を表示するとともに、選択情報付与工程により付与された選択情報を表示するので、多重に利用される文書の識別性および一度も選択されない文書の識別性を向上させることができ、これにより、任

意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０４１４】

また、請求項３６の発明によれば、請求項３１～３５の発明において、前記分類体系生成工程が、前記選択指示工程により選択されたクラスタ特徴のほか、前記文書の部分集合の中から選択された文書の部分集合に所属する文書群の全部あるいは一部および／または操作者が作成した情報に基づいて分類体系の構成要素を生成するので、クラスタの内容把握を容易にし、かつ、操作者独自の分類体系を簡易に生成できることので、分類体系の利用価値を向上させることができ、これにより、任意の文書集合にどのような内容が含まれるかを漸次的に収集することが可能な文書分類方法が得られるという効果を奏する。

【０４１５】

また、請求項３７の発明によれば、一つの文書の中に複数の話題や意味が含まれている場合に、ある特定の話題や意味に限定されたカテゴリに分類されたり、利用者の意図するカテゴリとは異なるカテゴリに分類されたりすることがなく、したがって、利用者がその分類カテゴリをよく理解できる。また、分割前文書（所属文書）中の分割文書の位置なども示されるので、利用者は文書群中の読みたい部分を効率的に読むことが可能な文書分類方法が得られるという効果を奏する。

【０４１６】

また、請求項３８の発明によれば、請求項２４～３７のいずれか一つに記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項２４～３７の動作をコンピュータによって実現することが可能な記録媒体が得られるという効果を奏する。

【図面の簡単な説明】

【図１】

この発明の実施の形態１による文書処理装置を構成する情報処理システム全体のハードウエア構成を示す説明図である。

【図２】

この発明の実施の形態１による文書処理装置を構成する情報処理システムにおけるサーバーのハードウエア構成を示す説明図である。

【図３】

この発明の実施の形態１による文書処理装置を構成する情報処理システムにおけるクライアントのハードウエア構成を示す説明図である。

【図４】

この発明の実施の形態１による文書処理装置の構成を機能的に示すブロック図である。

【図５】

この発明の実施の形態１による文書処理装置の項目名と項目値の関係を示す説明図である。

【図６】

この発明の実施の形態１による文書処理装置の文書記憶部に記憶された文書のデータ構造を示す説明図である。

【図７】

この発明の実施の形態１による文書処理装置の文書記憶部に記憶された文書の別のデータ構造を示す説明図である。

【図８】

この発明の実施の形態１による文書処理装置の出力部による画面表示の例を示す説明図である。

【図９】

この発明の実施の形態１による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図１０】

この発明の実施の形態１による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図１１】

この発明の実施の形態１による文書処理装置の特徴抽出部によりおこなわれる

抽出処理の内容の一覧を示す説明図である。

【図１２】

この発明の実施の形態１による文書処理装置の加工処理部によりおこなわれる加工処理の内容の一覧を示す説明図である。

【図１３】

この発明の実施の形態１による文書処理装置の各項目の特徴ベクトルを示す説明図である。

【図１４】

この発明の実施の形態１による文書処理装置の単語とその単語ＩＤごとの出現回数を示す説明図である。

【図１５】

この発明の実施の形態１による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図１６】

この発明の実施の形態１による文書処理装置の出力部によるクロス表作成のための指示画面を示す説明図である。

【図１７】

この発明の実施の形態１による文書処理装置の出力部による分類処理の結果が表示されたクロス表を示す説明図である。

【図１８】

この発明の実施の形態１による文書処理装置の出力部による分類処理の結果が表示された別のクロス表を示す説明図である。

【図１９】

この発明の実施の形態１による文書処理装置の出力部の詳細な構成を示すブロック図である。

【図２０】

この発明の実施の形態１による文書処理装置のクロス表の出力手順を示すフローチャートである。

【図２１】

この発明の実施の形態１による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図２２】

この発明の実施の形態１による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図２３】

この発明の実施の形態１による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図２４】

この発明の実施の形態１による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図２５】

この発明の実施の形態１による文書処理装置の文書記憶部の詳細な構成を示すブロック図である。

【図２６】

この発明の実施の形態１による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図２７】

この発明の実施の形態１による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図２８】

この発明の実施の形態１による文書処理装置の出力部による画面表示の別の例を示す説明図である。

【図２９】

この発明の実施の形態１による文書処理装置の文書処理の一連の手順を示すフローチャートである。

【図３０】

この発明の実施の形態２による文書分類装置の構成を機能的に示すブロック図である。

【図３１】

この発明の実施の形態２による文書分類装置のクラスタ特徴表示部の表示の一例を示す説明図である。

【図３２】

この発明の実施の形態２による文書分類装置の一連の処理の手順を示すフローチャートである。

【図３３】

この発明の実施の形態３による文書分類装置の構成を機能的に示すブロック図である。

【図３４】

この発明の実施の形態３よる文書分類装置の一連の処理の手順を示すフローチャートである。

【図３５】

この発明の実施の形態４による文書分類装置の構成を機能的に示すブロック図である。

【図３６】

この発明の実施の形態４よる文書分類装置の一連の処理の手順を示すフローチャートである。

【図３７】

この発明の実施の形態５による文書分類装置の構成を機能的に示すブロック図である。

【図３８】

この発明の実施の形態５よる文書分類装置の一連の処理の手順を示すフローチャートである。

【図３９】

この発明の実施の形態６による文書分類装置の構成を機能的に示すブロック図である。

【図４０】

この発明の実施の形態６による文書分類装置の分類結果記憶部において設けら

れたテーブルを示す説明図である。

【図４１】

この発明の実施の形態６による文書分類装置の選択情報付与部の処理の手順を示すフローチャートである。

【図４２】

この発明の実施の形態７を示す文書分類装置の構成ブロック図である。

【図４３】

この発明の実施の形態７による文書分類装置および文書分類方法の説明図である。

【図４４】

この発明の実施の形態７による文書分類装置および文書分類方法の他の説明図である。

【図４５】

この発明の実施の形態７による文書分類装置および文書分類方法の他の説明図である。

【図４６】

この発明の実施の形態７による文書分類装置および文書分類方法の他の説明図である。

【図４７】

この発明の実施の形態８による文書分類装置の構成ブロック図である。

【図４８】

この発明の実施の形態９による文書分類装置の構成ブロック図である。

【図４９】

この発明の実施の形態１０による文書分類装置および文書分類方法の説明図である。

【図５０】

この発明の実施の形態１１による文書分類装置および文書分類方法の説明図である。

【図５１】

この発明の実施の形態１２による文書分類装置の構成ブロック図である。

【図５２】

この発明の実施の形態１２による文書分類装置および文書分類方法の説明図である。

【図５３】

この発明の実施の形態１３による文書分類装置および文書分類方法の説明図である。

【図５４】

この発明の実施の形態１４による文書分類装置および文書分類方法の説明図である。

【図５５】

この発明の実施の形態１５による文書分類装置および文書分類方法の説明図である。

【図５６】

この発明の実施の形態１６による文書分類装置および文書分類方法の説明図である。

【符号の説明】

１０１　サーバー

１０２　クライアント

１０３　ネットワーク

２０１　ＣＰＵ

２０４　Ｉ／Ｆ

２０６　ディスク装置

３０１　ＣＰＵ

３０６　ハードディスク

３０８　ディスプレイ

３０９　Ｉ／Ｆ

３１１　キーボード

３１２　マウス

３１３　スキャナ

４００　バス

４０１　入力部

４０２　文書記憶部

４０３　選択部

４０４　特徴抽出部

４０５　加工処理部

４０６　出力部

４０７　グラフ描画部

４０８　加工処理結果保持部

４０９　解析部

４１０　特徴ベクトル生成部

１６００　指示画面

１７００　クロス表

１８００　マウスポインタ

１８０１　内容表示画面

１９０１　項目値選定部

１９０２　集計部

１９０３　表保持部

２４０１　棒グラフ表示領域

２５０１　設定値記憶部

２５０２　設定値送受信部

２５０３　分類情報記憶部

２６０４　問い合わせ画面

２７０３　分類情報表示画面

２８０２　表示領域

３００１　入力部

３００２　言語解析部

３００３　ベクトル生成部

３００４　分類部

３００５　分類パラメータ指示部

３００６　分類結果記憶部

３００７　クラスタ特徴表示部

３００８　クラスタ特徴算出部

３００９　分類体系記憶部

３０１０　クラスタ選択指示部

３０１１　分類体系閲覧操作部

３１１０　カーソル

３３０１，３５０１，３７０１　ベクトル記憶部

３３０２，３７０２　ベクトル修正部

３５０２，３７０３　文書表現空間修正部

３９０１　選択情報付与部

４０００　テーブル

５００１　文書入力部

５００２　文書分割部

５００３　文書－分割文書対応マップ生成部

５００４　分割文書分類部

５００５　分割文書分類結果生成部

５００６　文書分類結果生成部

５００７　文書保存部

５００８　分割文書保存部

５００９　文書－分割文書対応マップ保存部

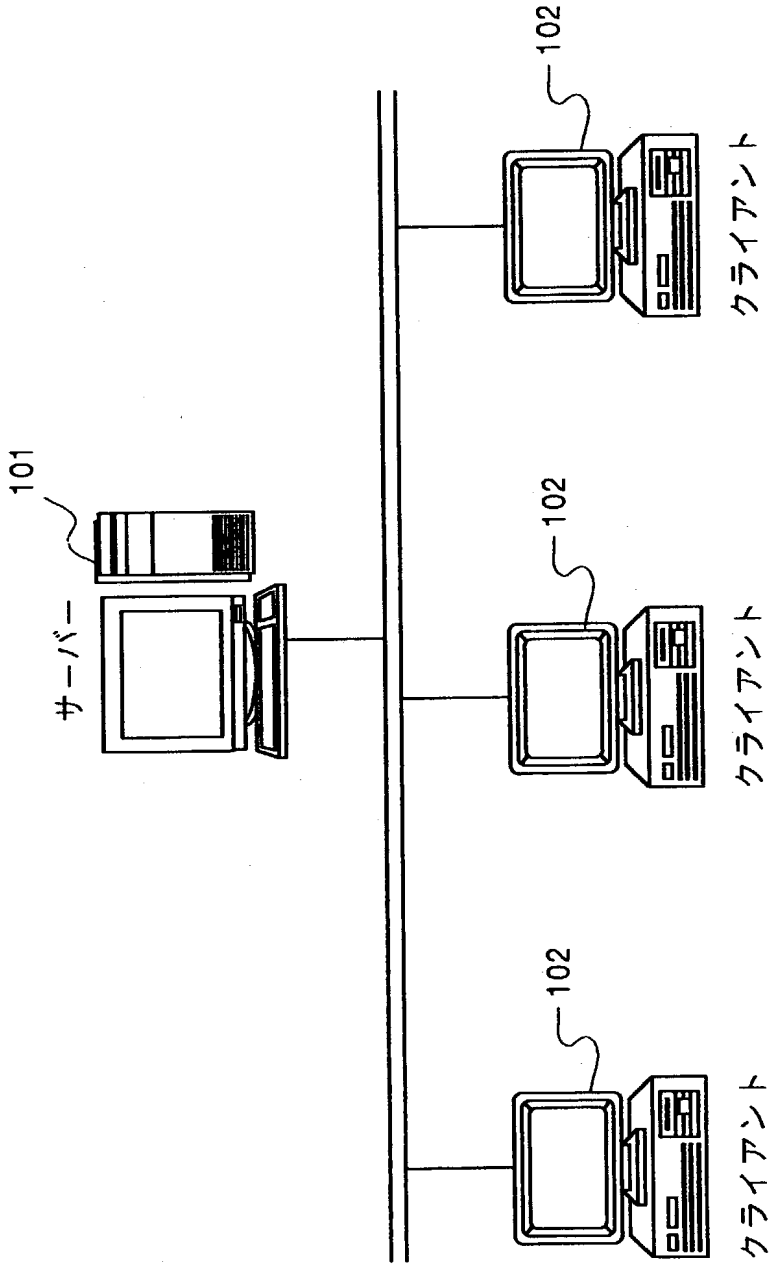
５０１０　分割文書分類結果保存部

５０１１　文書要素解析部

５０１２　要素付随情報抽出部

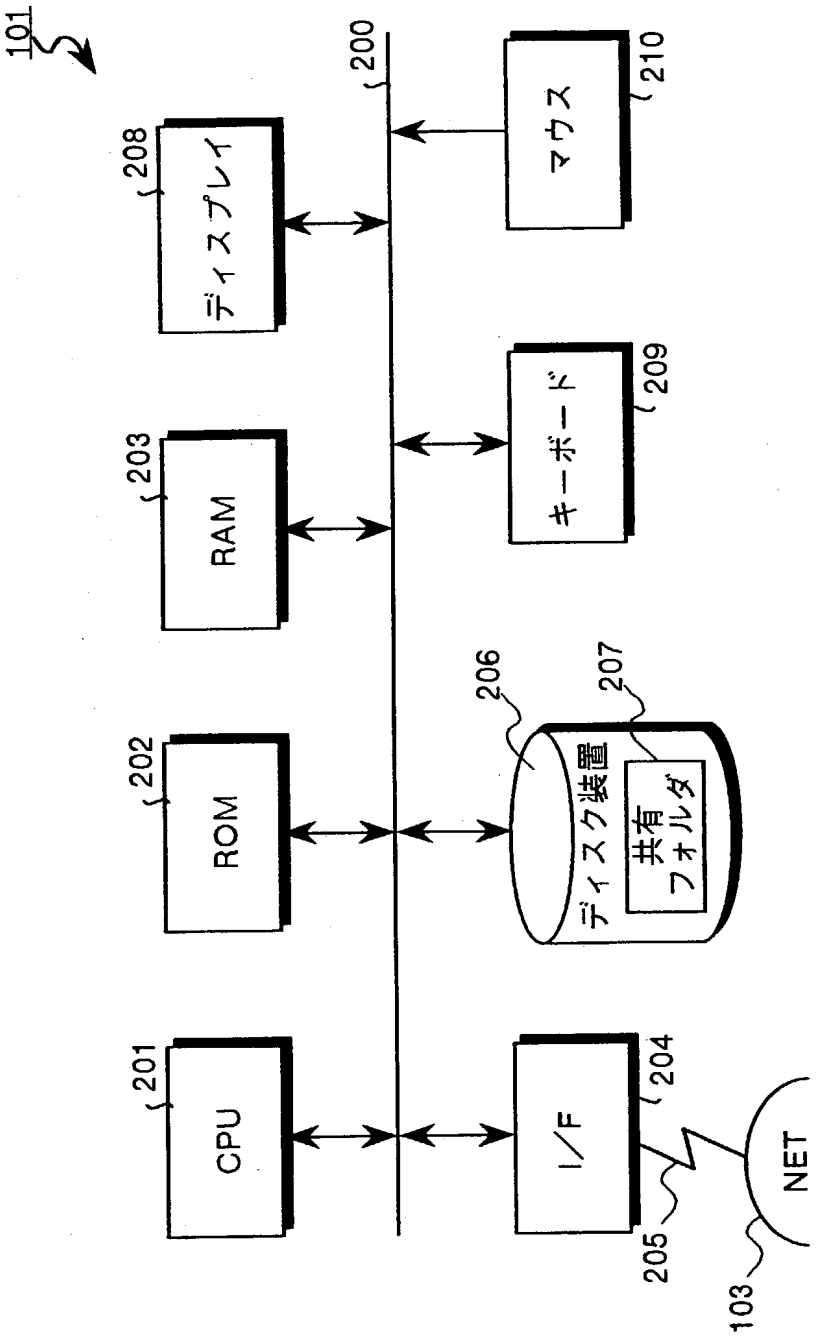
【書類名】　　図面

【図１】

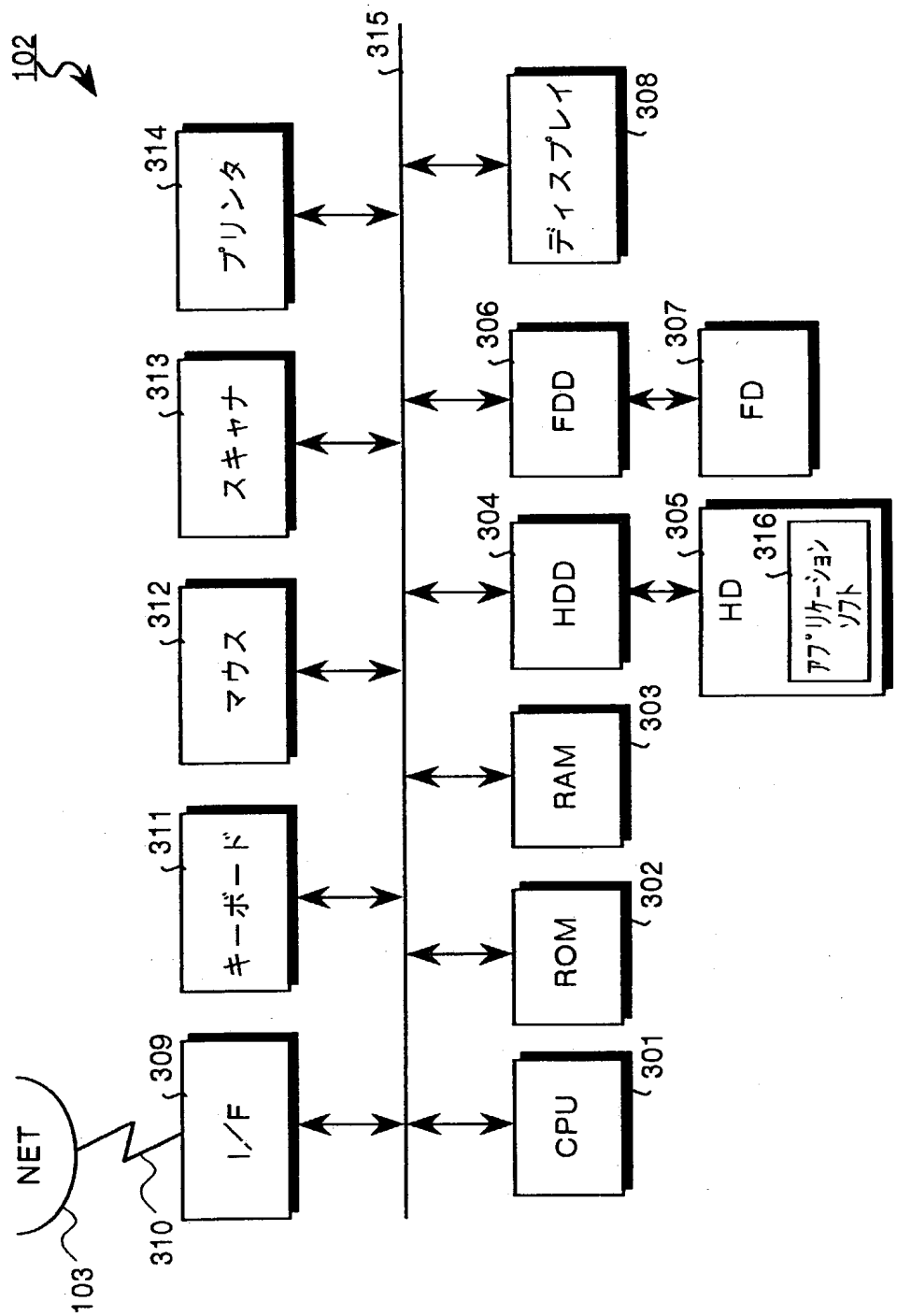


出証特平１１－３０９１７２６

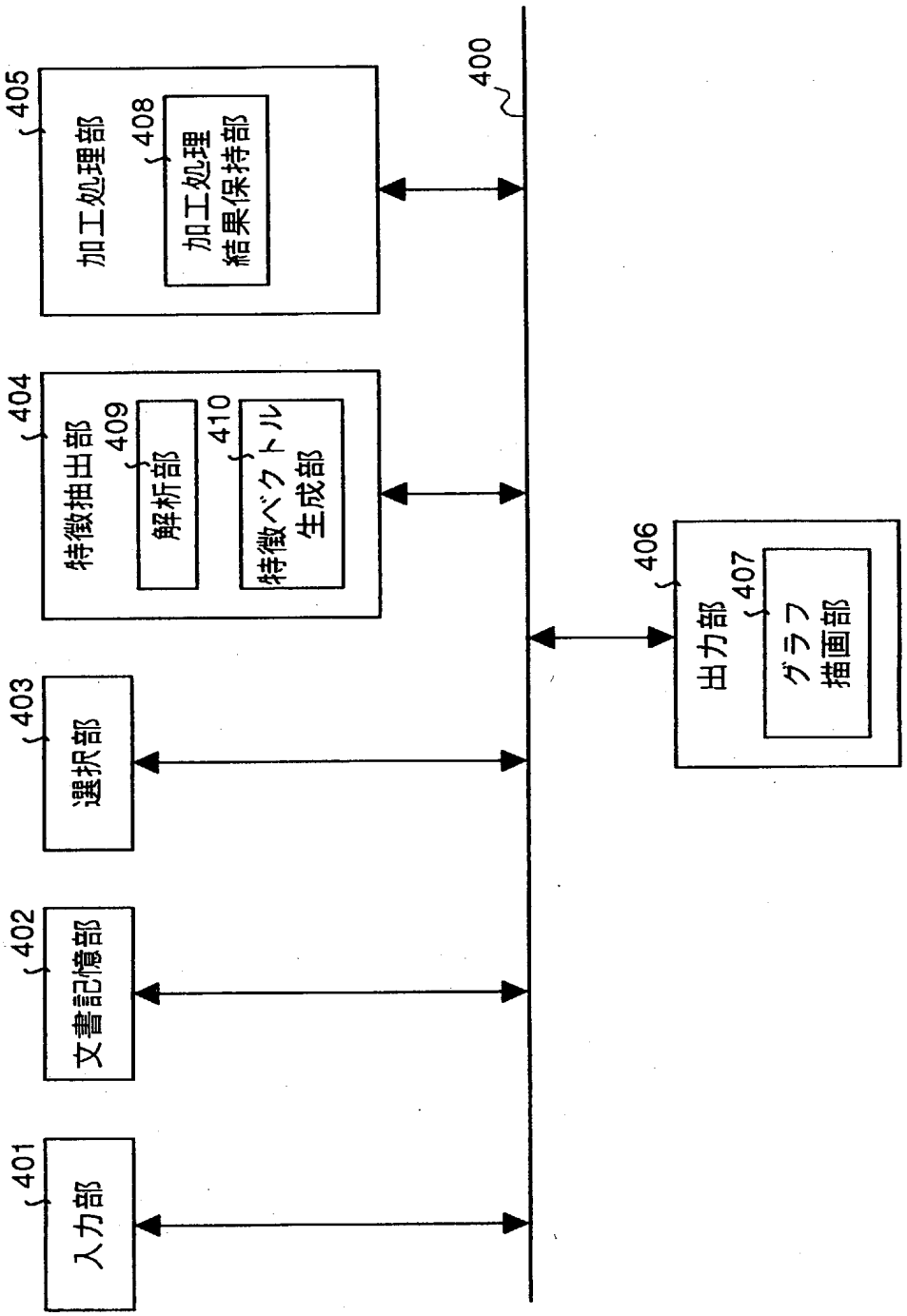
【図２】


101
200
208
ディスプレイ
210
マウス
209
キーボード
203
RAM
206
207
202
ROM
ディスク装置
共有
フォルダ
201
CPU
204
I/F
NET
205
103

【図３】


102
315
314
プリンタ
308
ディスプレイ
313
スキャナ
306
FDD
307
FD
312
マウス
304
HDD
305
HD
316
アプリケーション
ソフト
311
キーボード
303
RAM
309
I/F
302
ROM
NET
310
301
CPU
103

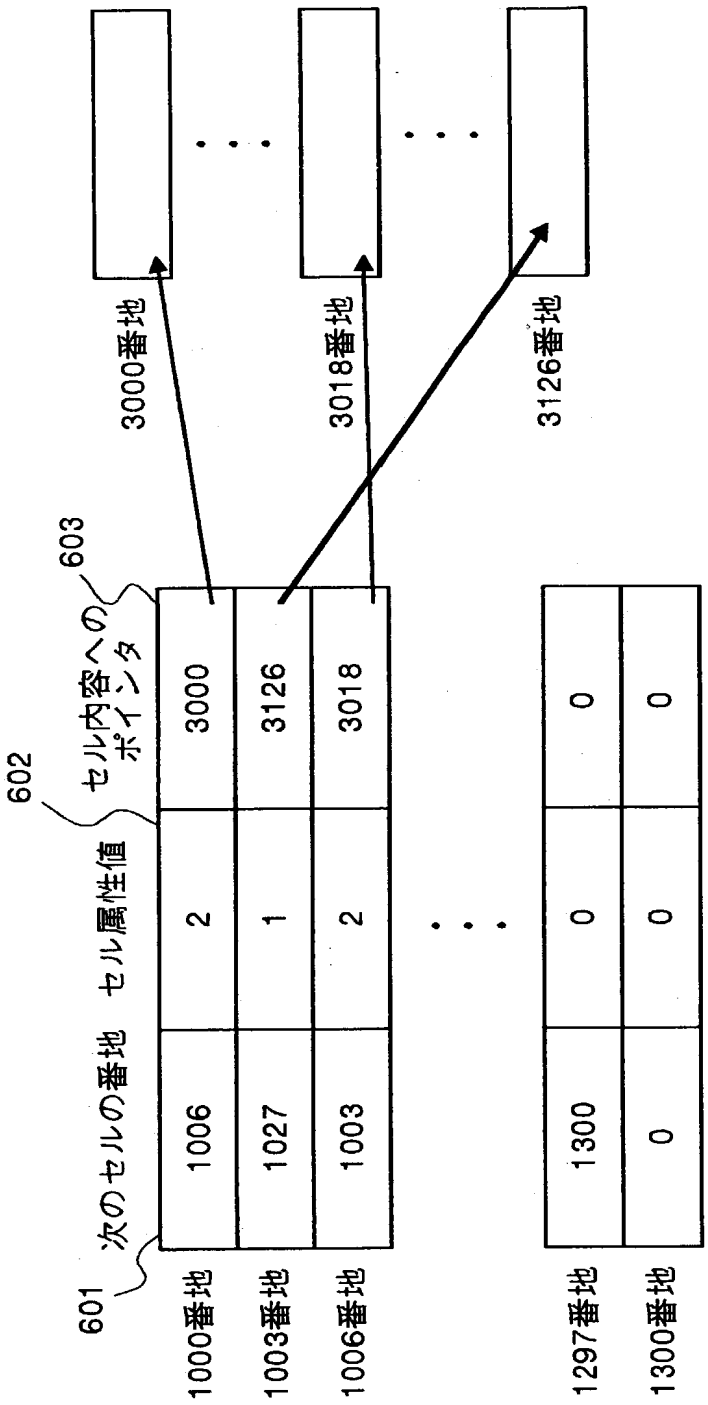
【図４】


400
405
加工処理部
408
加工処理
結果保持部
404
特徴抽出部
409
解析部
410
特徴ベクトル
生成部
406
出力部
407
グラフ
描画部
403
選択部
402
文書記憶部
401
入力部

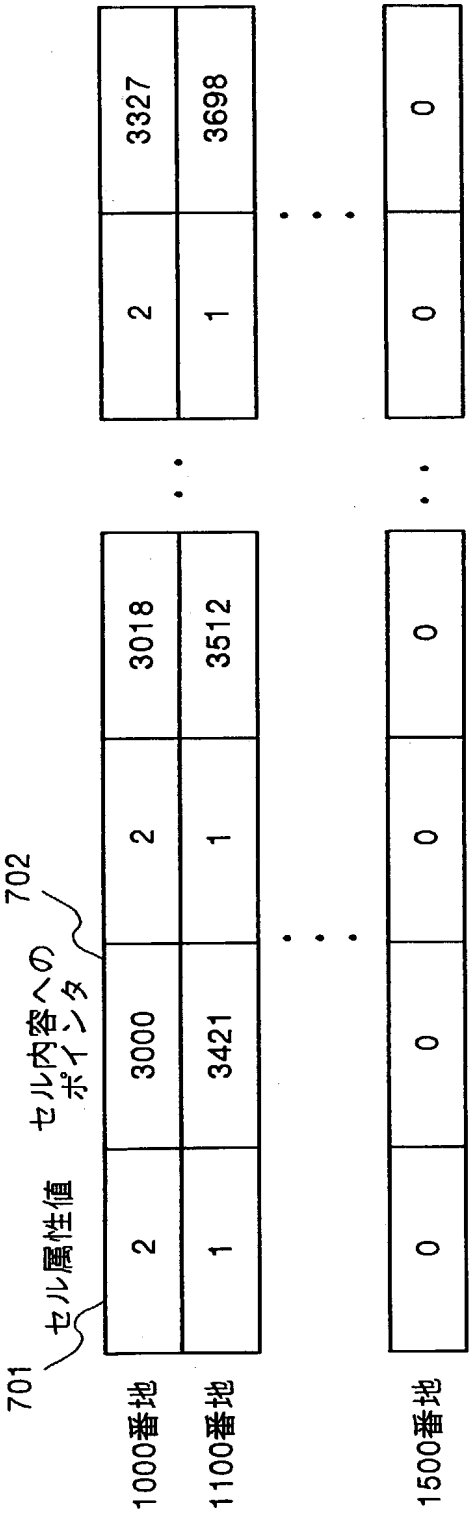
【図５】

項目名	項目値
出願番号	特願平１０－０００００
出願日	平成１０年１月１日
請求項の数	3
発明の名称	画像形成装置
発明の効果	本発明の画像形成装置は、○○を××しているので…

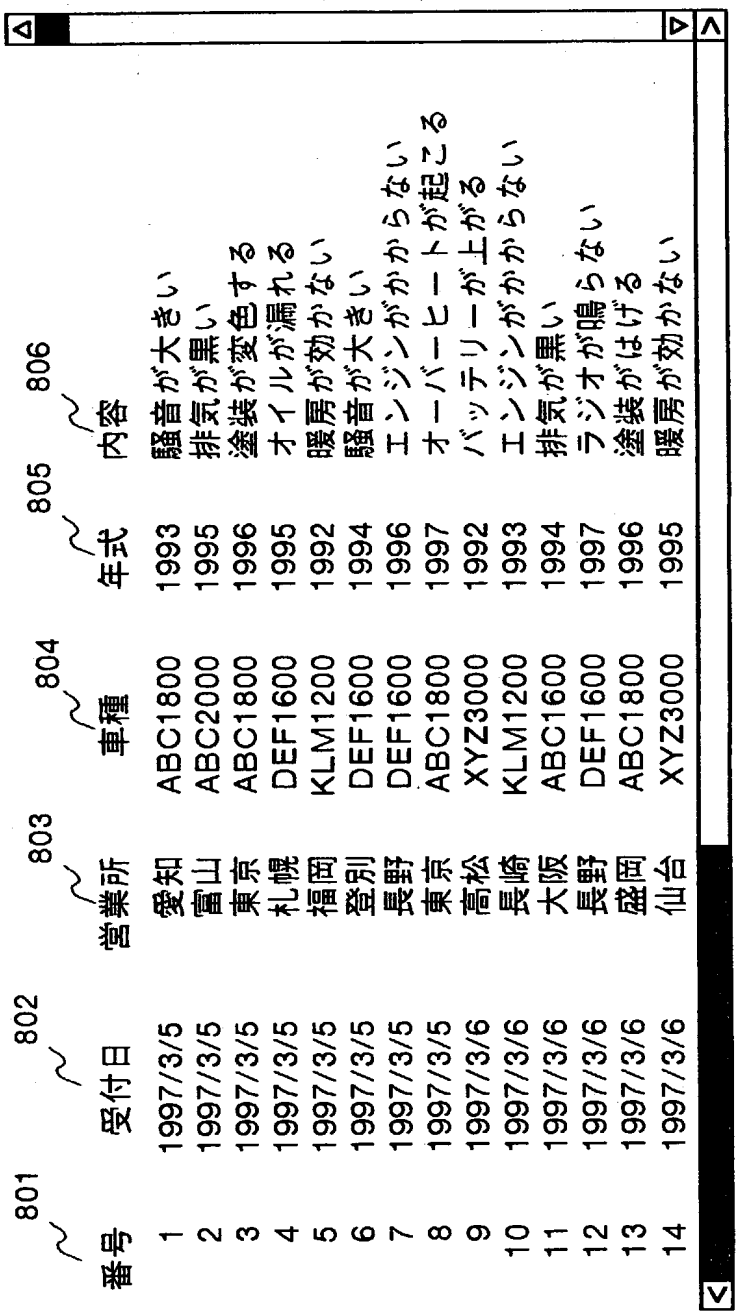
【図６】


3000番地
3018番地
3126番地
603
602
601
次のセルの番地
セル属性値
セル内容へのポインタ
1006
2
3000
1027
1
3126
1003
2
3018
1300
0
0
0
0
0
1000番地
1003番地
1006番地
1297番地
1300番地

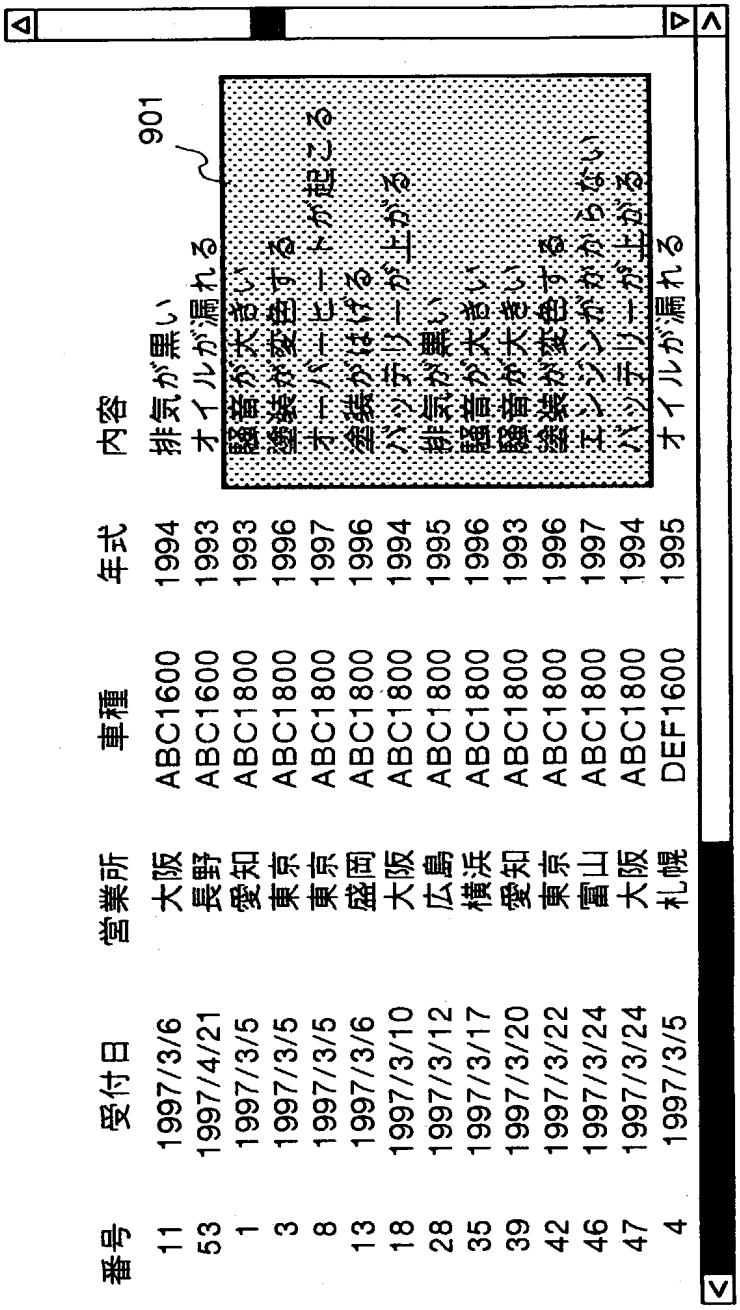
【図７】


701
セル属性値
セル内容へのポインタ
702
1000番地
1100番地
1500番地
2
3000
3018
3327
1
3421
3512
3698
0
0
0
0
0
0
0
0

【図８】


801
802
803
804
805
806
番号
受付日
営業所
車種
年式
内容
1 1997/3/5 愛知 ABC1800 1993 騒音が大きい
2 1997/3/5 富山 ABC2000 1995 排気が黒い
3 1997/3/5 東京 ABC1800 1996 塗装が変色する
4 1997/3/5 札幌 DEF1600 1995 オイルが漏れる
5 1997/3/5 福岡 KLM1200 1992 暖房が効かない
6 1997/3/5 登別 DEF1600 1994 騒音が大きい
7 1997/3/5 長野 DEF1600 1996 エンジンがかからない
8 1997/3/5 東京 ABC1800 1997 オーバーヒートが起こる
9 1997/3/6 高松 XYZ3000 1992 バッテリーが上がる
10 1997/3/6 長崎 KLM1200 1993 エンジンがかからない
11 1997/3/6 大阪 ABC1600 1994 排気が黒い
12 1997/3/6 長野 DEF1600 1997 ラジオが鳴らない
13 1997/3/6 盛岡 ABC1800 1996 塗装がはげる
14 1997/3/6 仙台 XYZ3000 1995 暖房が効かない

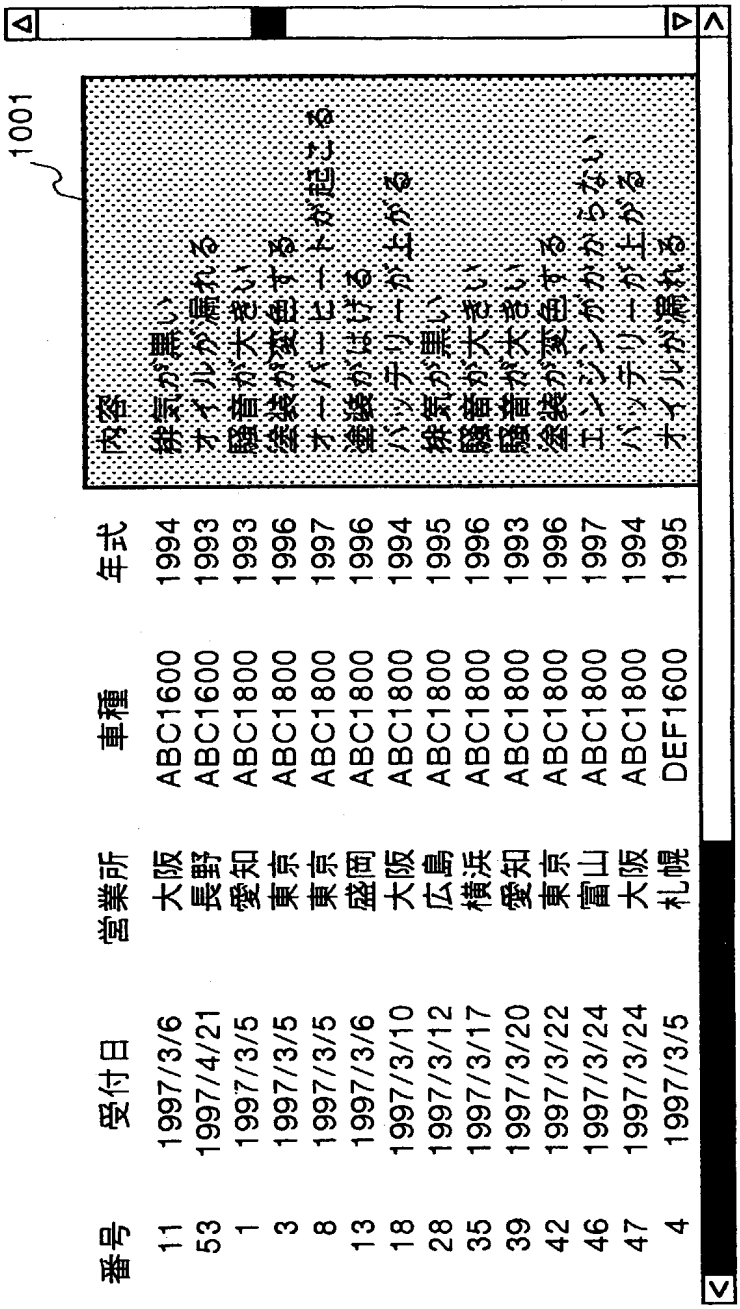
【図９】



番号	受付日	営業所	車種	年式	内容
11	1997/3/6	大阪	ABC1600	1994	排気が黒い
53	1997/4/21	長野	ABC1600	1993	オイルが漏れる
1	1997/3/5	愛知	ABC1800	1993	騒音が大きい
3	1997/3/5	東京	ABC1800	1996	塗装が変色する
8	1997/3/5	東京	ABC1800	1997	オイルが減る
13	1997/3/6	盛岡	ABC1800	1996	塗装がはがれる
18	1997/3/10	大阪	ABC1800	1994	バッテリーが上がる
28	1997/3/12	広島	ABC1800	1995	排気が黒い
35	1997/3/17	横浜	ABC1800	1996	騒音が大きい
39	1997/3/20	愛知	ABC1800	1993	騒音が大きい
42	1997/3/22	東京	ABC1800	1996	塗装が変色する
46	1997/3/24	富山	ABC1800	1997	エンジンがかからない
47	1997/3/24	大阪	ABC1800	1994	バッテリーが上がる
4	1997/3/5	札幌	DEF1600	1995	オイルが漏れる

901

【図１０】


1001
内容
排気が黒い
オイルが漏れる
騒音が大きい
燃費が悪くなる
オーバーヒートが起こる
異音がする
エンジンがかからない
ハンドルが重い
オイルが漏れる
年式
1994
1993
1993
1996
1997
1995
1996
1994
1995
1996
1997
1994
1995
車種
ABC1600
ABC1600
ABC1800
ABC1800
ABC1800
ABC1800
ABC1800
ABC1800
ABC1800
ABC1800
ABC1800
ABC1800
DEF1600
営業所
大阪
長野
愛知
東京
東京
盛岡
大阪
広島
横浜
愛知
東京
富山
大阪
札幌
受付日
1997/3/6
1997/4/21
1997/3/5
1997/3/5
1997/3/5
1997/3/6
1997/3/10
1997/3/12
1997/3/17
1997/3/20
1997/3/22
1997/3/24
1997/3/24
1997/3/5
番号
11
53
1
3
8
13
18
28
35
39
42
46
47
4

【図１１】

	抽出処理内容
1	対象とする文字列に含まれる単語
2	対象とする文字列に含まれる単語数
3	対象とする文字列に含まれる単語の文字数
4	対象とする文字列に含まれる単語それぞれの出現回数
5	対象とする文字列に含まれる単語それぞれの品詞
6	対象とする文字列に含まれる単語間の関係の情報
7	対象とする文字列に含まれる文の数
8	対象とする文字列に含まれる文の文字数
9	対象とする文字列に含まれる文の文節数
10	対象とする文字列に含まれる文の間の関係
…	…

【図１２】

	加工処理内容
1	分類処理
2	検索処理
3	並べ替え処理
4	代表値抽出処理
5	最大値抽出処理
6	最小値抽出処理
7	算術処理
⋮	⋮

出証特平１１－３０９１７２６

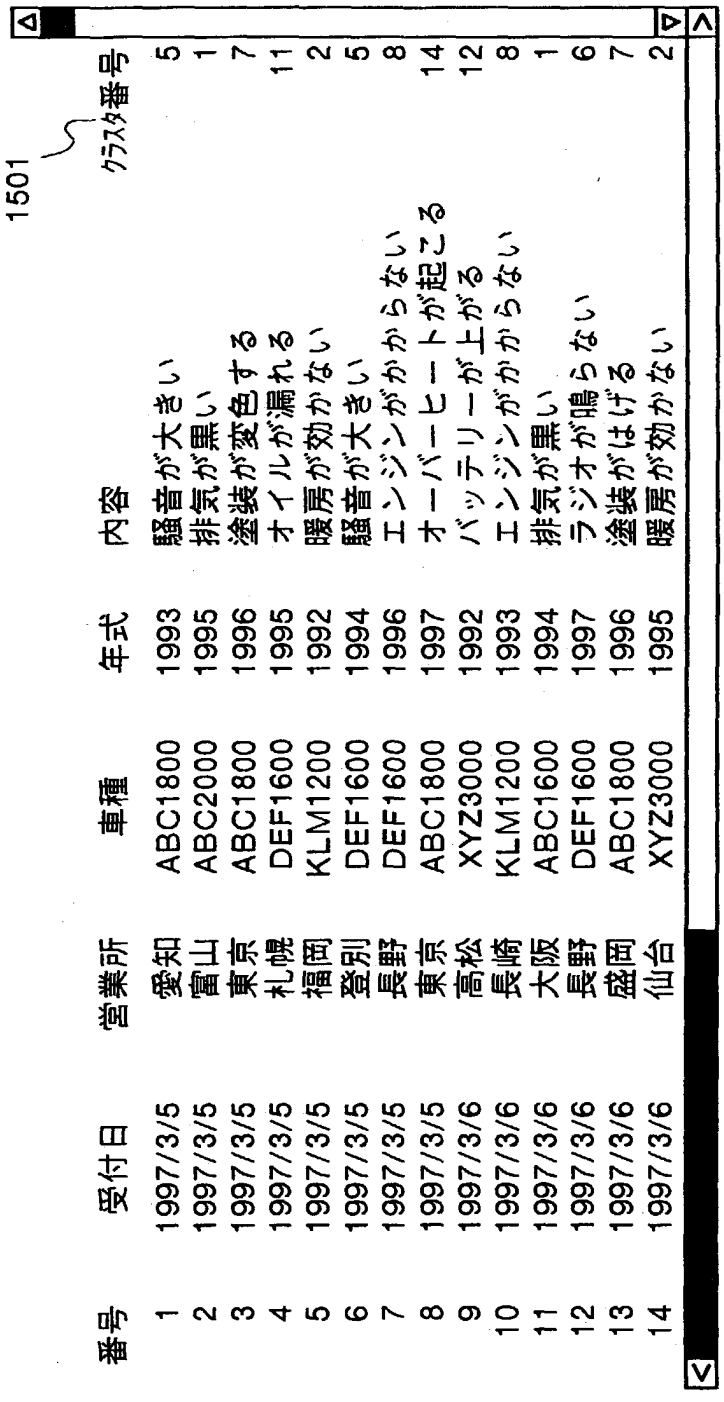
【図１３】

	騒音	が	大きい	塗装	変色	する	オーバーヒート	起こる	はげる	バッテリー	上がる	排気	黒い
騒音が大きい	1	1	1	0	0	0	0	0	0	0	0	0	0
塗装が変色する	0	1	0	1	1	1	0	0	0	0	0	0	0
オーバーヒートが起こる	0	1	0	0	0	0	1	1	0	0	0	0	0
塗装がはげる	0	1	0	1	0	0	0	0	1	0	0	0	0
バッテリーが上がる	0	1	0	0	0	0	0	0	0	1	1	0	0
排気が黒い	0	1	0	0	0	0	0	0	0	0	0	1	1

【図１４】

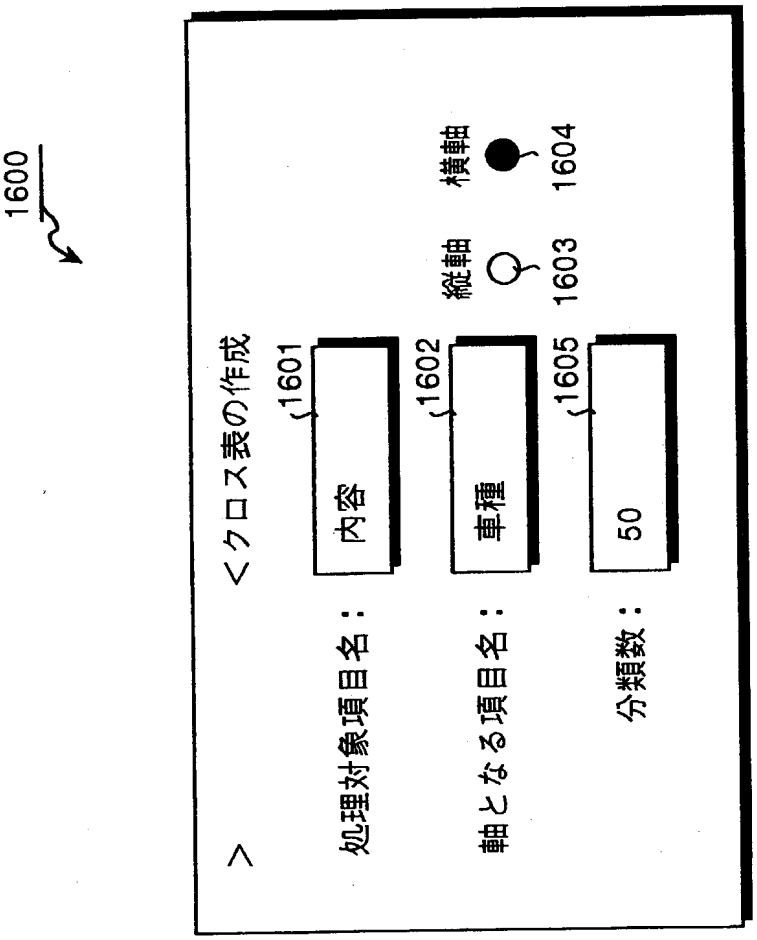
単語	単語ID
騒音	1
が	2
大きい	3
塗装	4
変色	5
する	6
ﾋｰﾄ	7
起こる	8
はげる	9
ﾊﾞｯﾃﾘｰ	10
上がる	11
排気	12
黒い	13

【図１５】

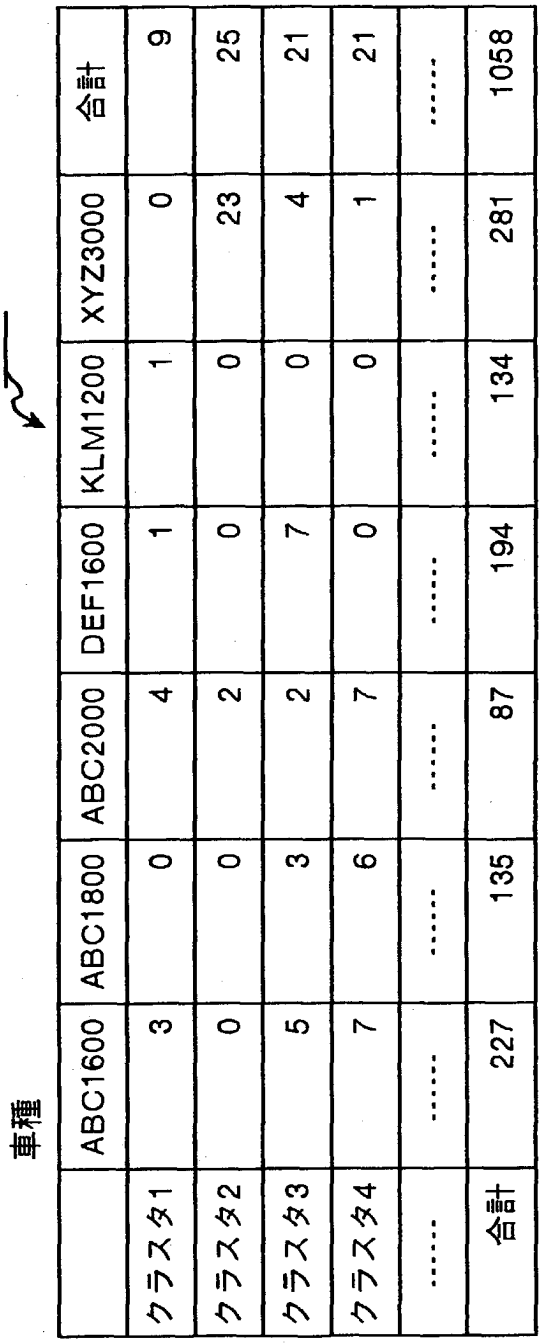


番号	受付日	営業所	車種	年式	内容	クラスタ番号
1	1997/3/5	愛知	ABC1800	1993	騒音が大きい	5
2	1997/3/5	富山	ABC2000	1995	排気が黒い	1
3	1997/3/5	東京	ABC1800	1996	塗装が変色する	7
4	1997/3/5	札幌	DEF1600	1995	オイルが漏れる	11
5	1997/3/5	福岡	KLM1200	1992	暖房が効かない	2
6	1997/3/5	登別	DEF1600	1994	騒音が大きい	5
7	1997/3/5	長野	DEF1600	1996	エンジンがかからない	8
8	1997/3/5	東京	ABC1800	1997	オーバーヒートが起こる	14
9	1997/3/6	高松	XYZ3000	1992	バッテリーが上がる	12
10	1997/3/6	長崎	KLM1200	1993	エンジンがかからない	8
11	1997/3/6	大阪	ABC1600	1994	排気が黒い	1
12	1997/3/6	長野	DEF1600	1997	ラジオが鳴らない	6
13	1997/3/6	盛岡	ABC1800	1996	塗装がはげる	7
14	1997/3/6	仙台	XYZ3000	1995	暖房が効かない	2

【図１６】

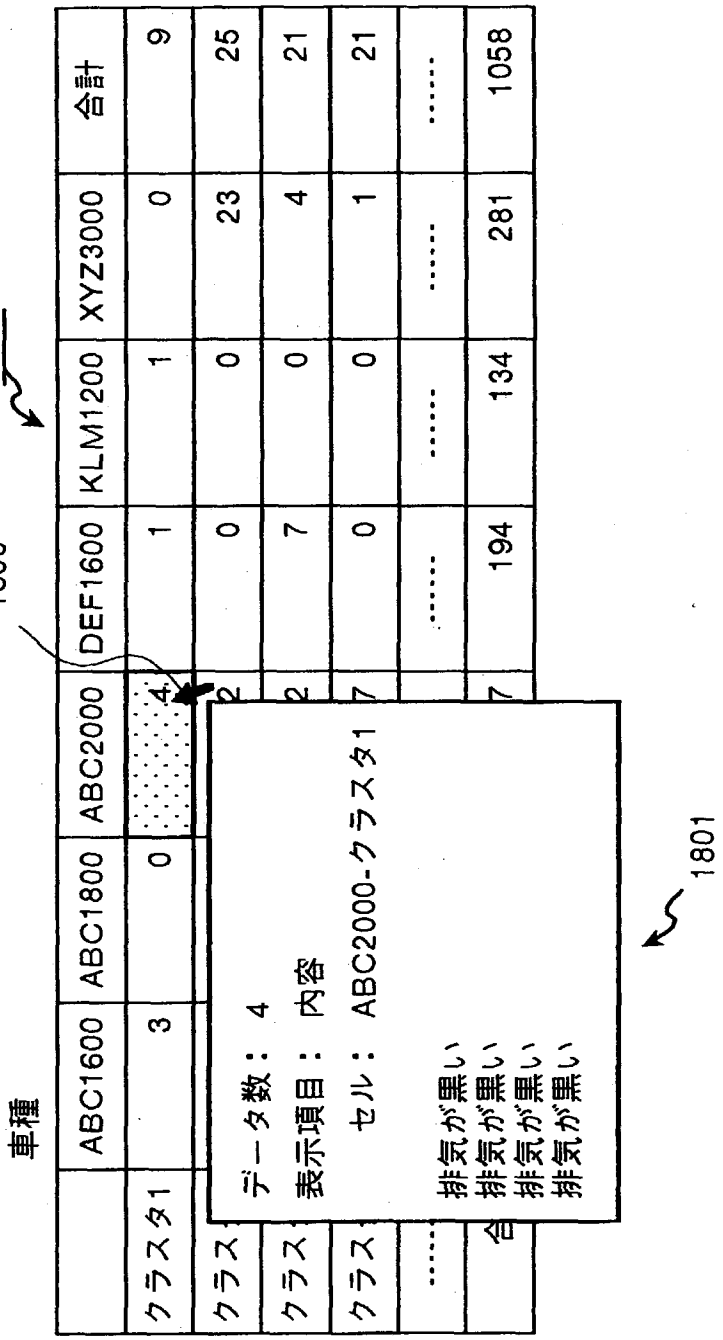

1600
＜クロス表の作成
処理対象項目名：
内容
1601
軸となる項目名：
車種
1602
縦軸
1603
横軸
1604
分類数：
50
1605
＞

【図１７】

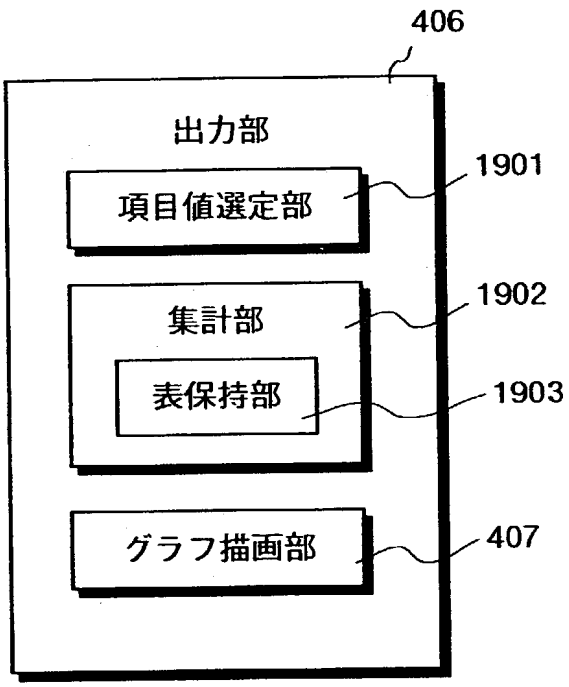


	ABC1600	ABC1800	ABC2000	DEF1600	KLM1200	XYZ3000	合計
クラスタ1	3	0	4	1	1	0	9
クラスタ2	0	0	2	0	0	23	25
クラスタ3	5	3	2	7	0	4	21
クラスタ4	7	6	7	0	0	1	21
.....	.....	.....	.....	.....	.....	.....	.....
合計	227	135	87	194	134	281	1058

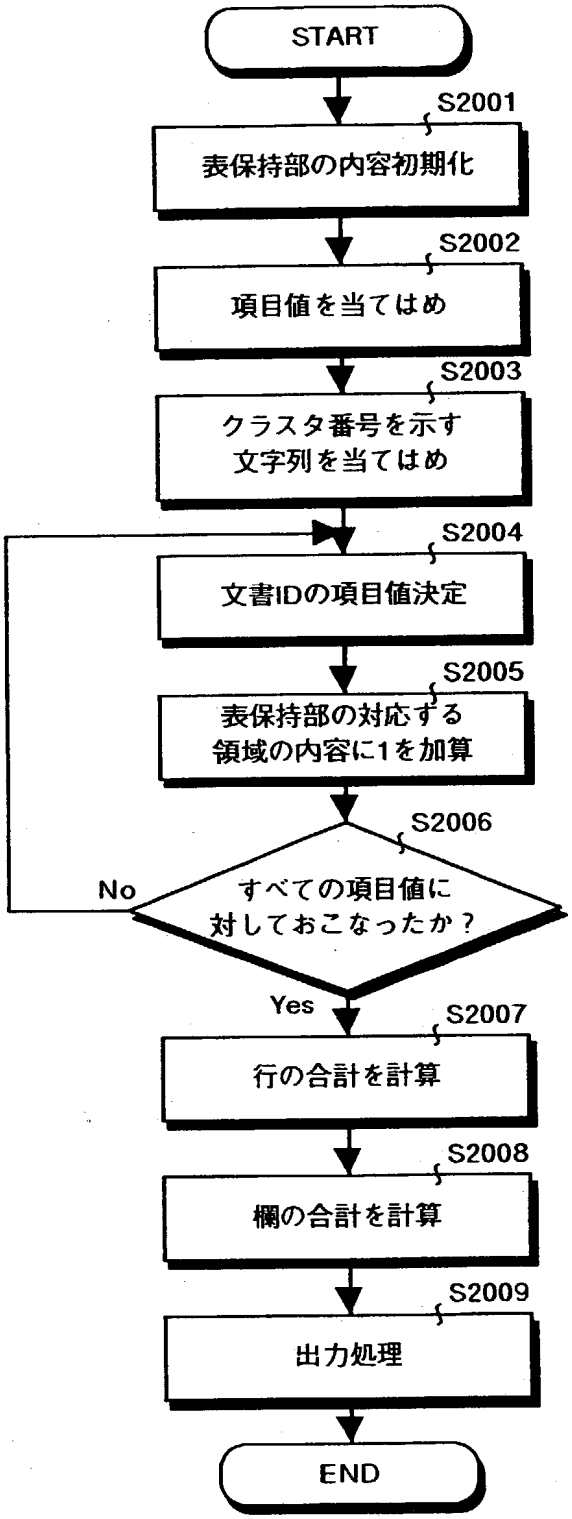
【図１８】


1700
1800
1801
車種
ABC1600
ABC1800
ABC2000
DEF1600
KLM1200
XYZ3000
合計
クラスタ1
3
0
4
0
1
0
9
クラス
2
0
0
23
25
クラス
2
7
0
4
21
クラス
7
0
0
1
21
.....
.....
.....
.....
.....
.....
.....
合
7
194
134
281
1058
データ数： 4
表示項目： 内容
セル： ABC2000-クラスタ1
排気が黒い
排気が黒い
排気が黒い
排気が黒い

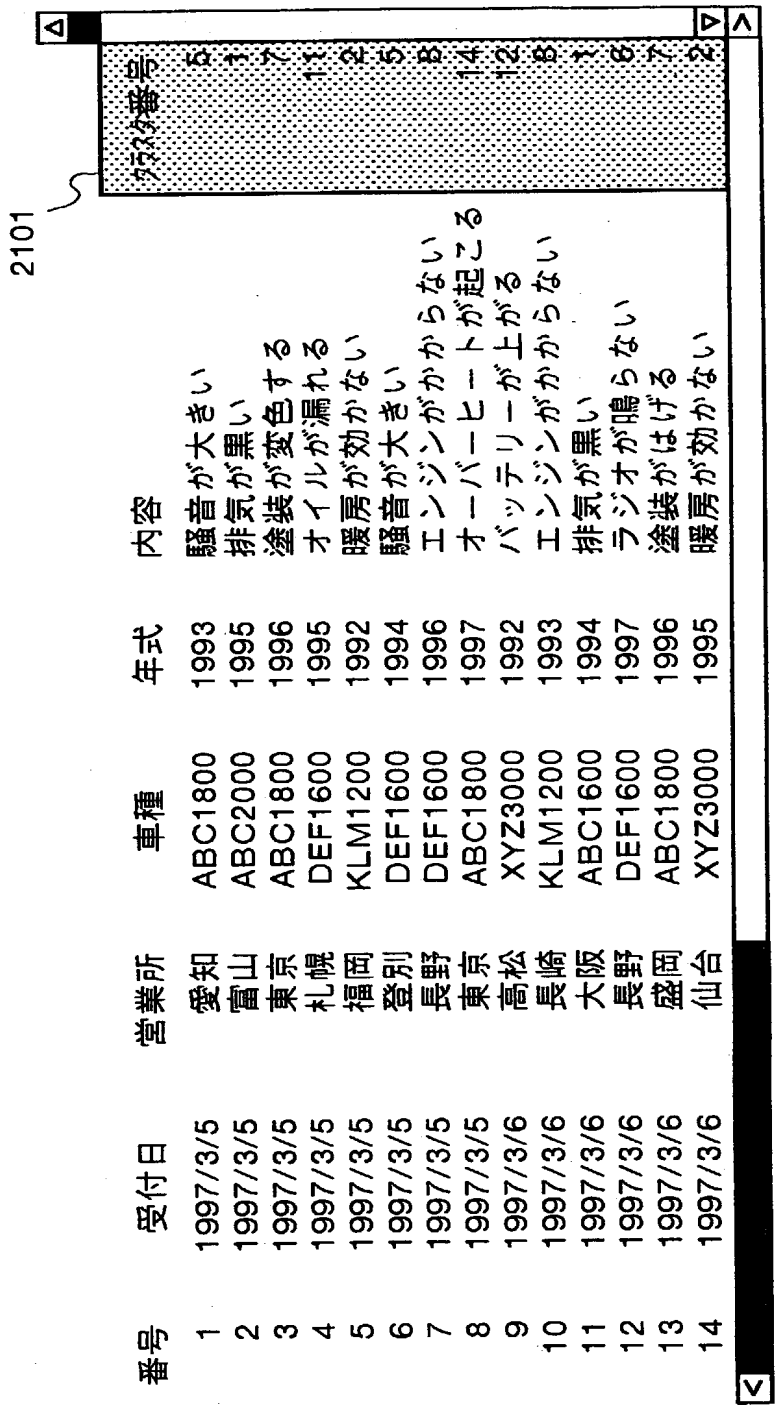
【図１９】



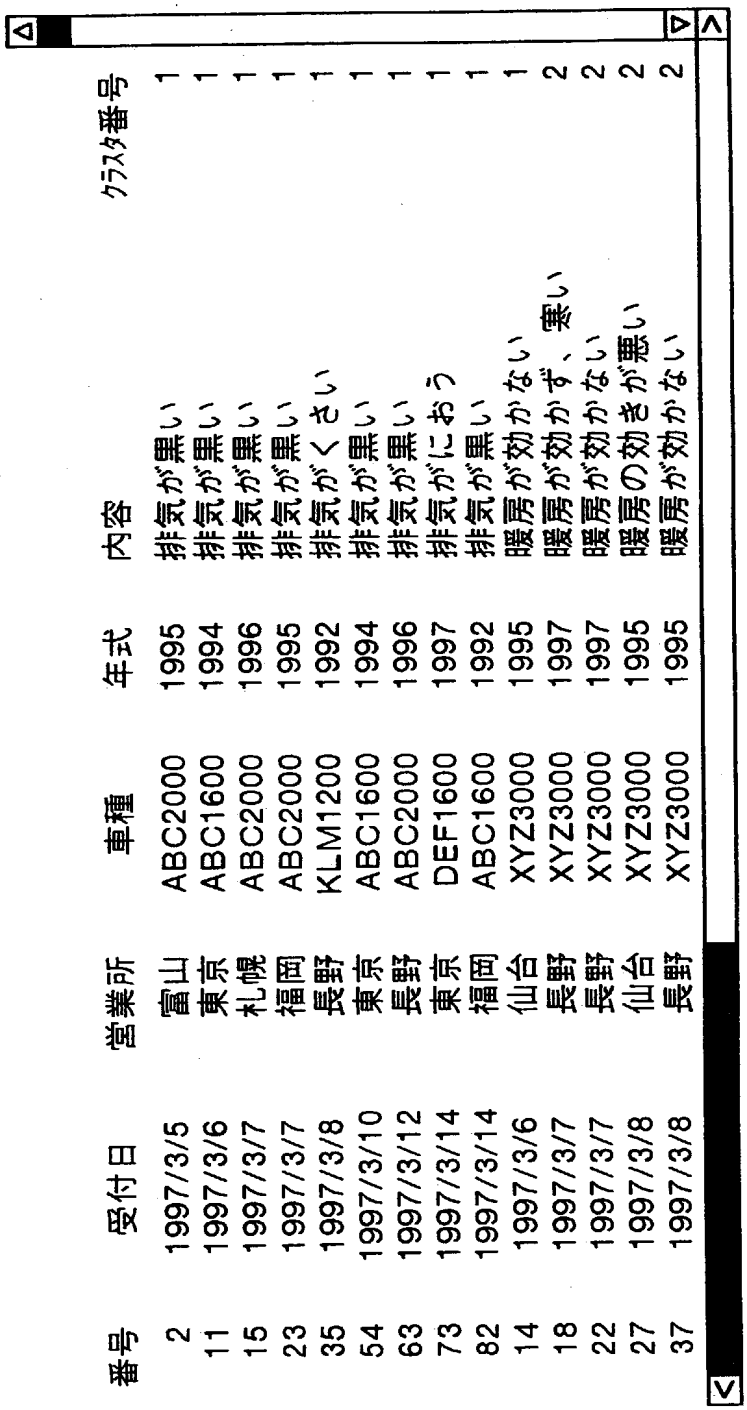
【図２０】


START
S2001
表保持部の内容初期化
S2002
項目値を当てはめ
S2003
クラスタ番号を示す
文字列を当てはめ
S2004
文書IDの項目値決定
S2005
表保持部の対応する
領域の内容に1を加算
S2006
No
すべての項目値に
対しておこなったか？
Yes
S2007
行の合計を計算
S2008
欄の合計を計算
S2009
出力処理
END

【図２１】

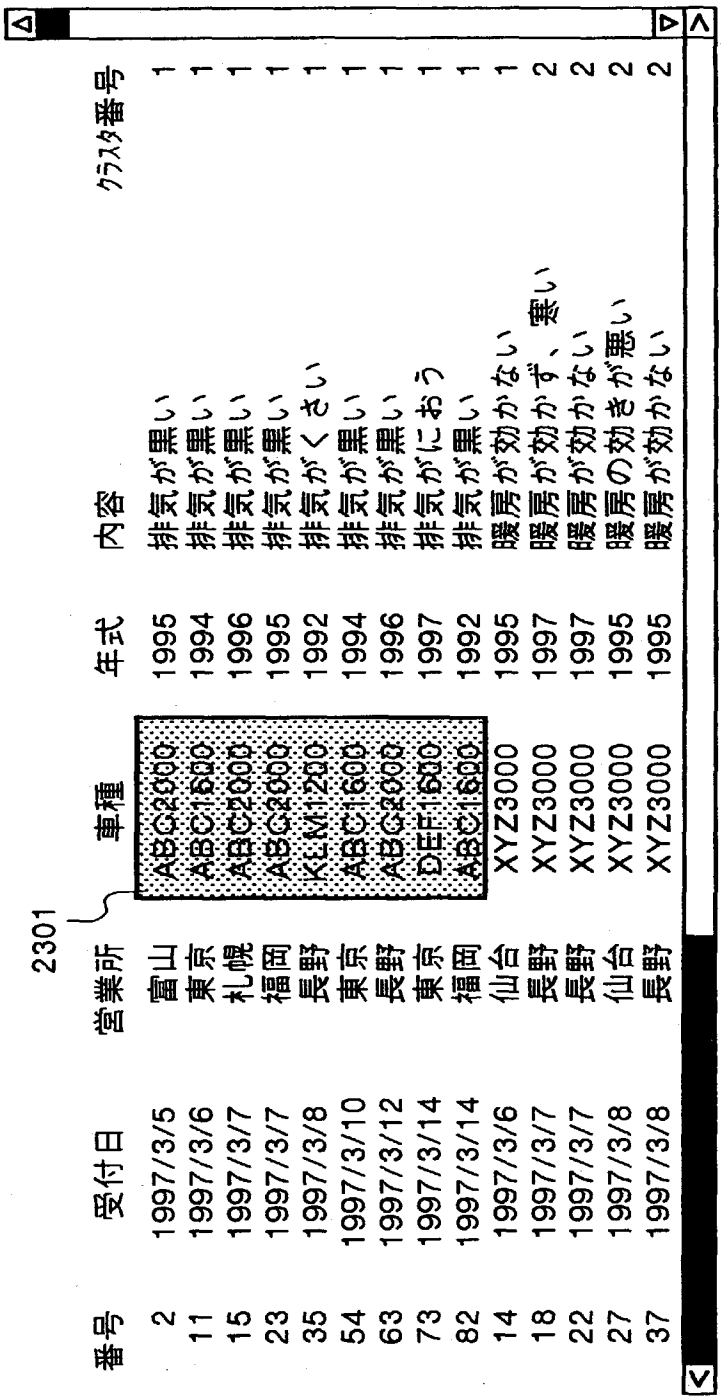

2101
クラス番号 5 1 7 1～2 5 8 1～4 1～2 8 1～6 7～2
番号 受付日 営業所 車種 年式 内容
1 1997/3/5 愛知 ABC1800 1993 騒音が大きい
2 1997/3/5 富山 ABC2000 1995 排気が黒い
3 1997/3/5 東京 ABC1800 1996 塗装が変色する
4 1997/3/5 札幌 DEF1600 1995 オイルが漏れる
5 1997/3/5 福岡 KLM1200 1992 暖房が効かない
6 1997/3/5 登別 DEF1600 1994 騒音が大きい
7 1997/3/5 長野 DEF1600 1996 エンジンがかからない
8 1997/3/5 東京 ABC1800 1997 オーバーヒートが起こる
9 1997/3/6 高松 XYZ3000 1992 バッテリーが上がる
10 1997/3/6 長崎 KLM1200 1993 エンジンがかからない
11 1997/3/6 大阪 ABC1600 1994 排気が黒い
12 1997/3/6 長野 DEF1600 1997 ラジオが鳴らない
13 1997/3/6 盛岡 ABC1800 1996 塗装がはげる
14 1997/3/6 仙台 XYZ3000 1995 暖房が効かない

【図２２】

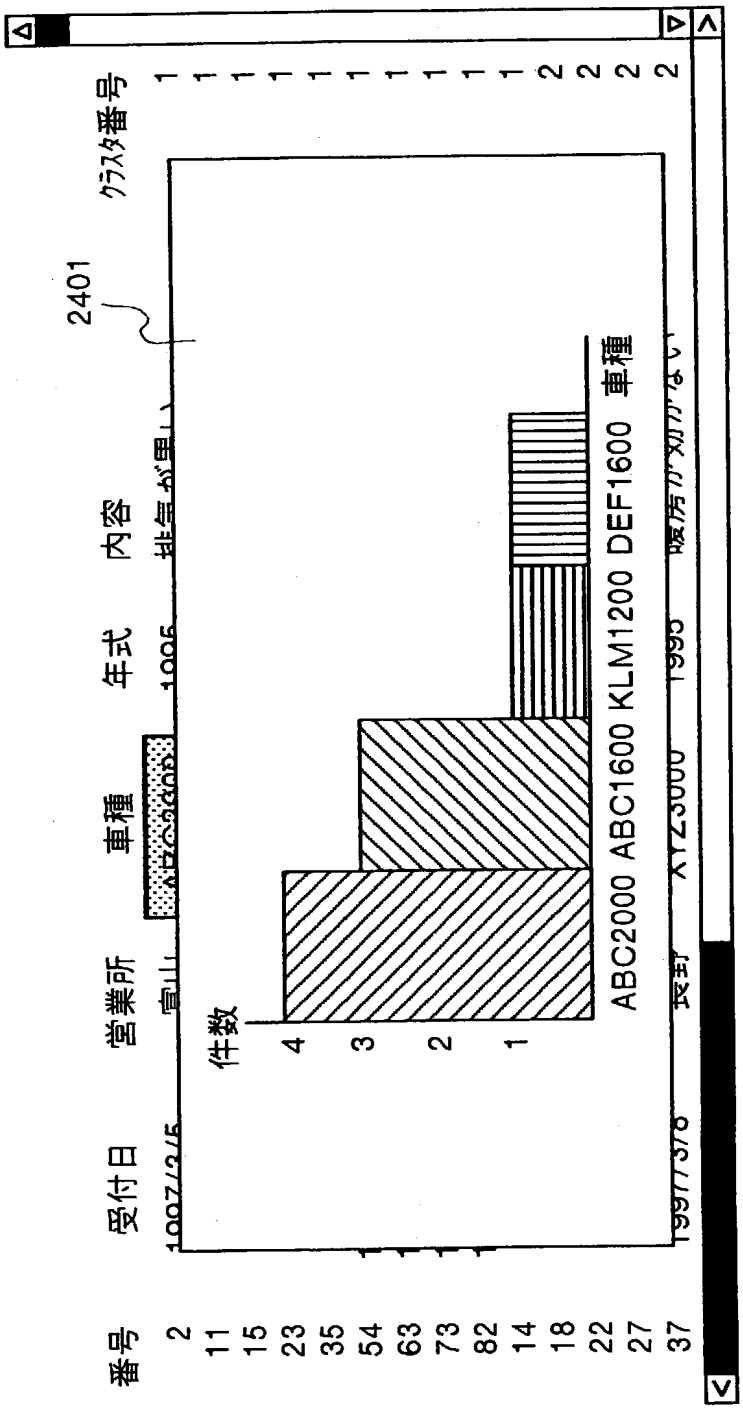


番号	受付日	営業所	車種	年式	内容	クラスタ番号
2	1997/3/5	富山	ABC2000	1995	排気が黒い	1
11	1997/3/6	東京	ABC1600	1994	排気が黒い	1
15	1997/3/7	札幌	ABC2000	1996	排気が黒い	1
23	1997/3/7	福岡	ABC2000	1995	排気が黒い	1
35	1997/3/8	長野	KLM1200	1992	排気がくさい	1
54	1997/3/10	東京	ABC1600	1994	排気が黒い	1
63	1997/3/12	長野	ABC2000	1996	排気が黒い	1
73	1997/3/14	東京	DEF1600	1997	排気がにおう	1
82	1997/3/14	福岡	ABC1600	1992	排気が黒い	1
14	1997/3/6	仙台	XYZ3000	1995	暖房が効かない	1
18	1997/3/7	長野	XYZ3000	1997	暖房が効かず、寒い	2
22	1997/3/7	長野	XYZ3000	1997	暖房が効かない	2
27	1997/3/8	仙台	XYZ3000	1995	暖房の効きが悪い	2
37	1997/3/8	長野	XYZ3000	1995	暖房が効かない	2

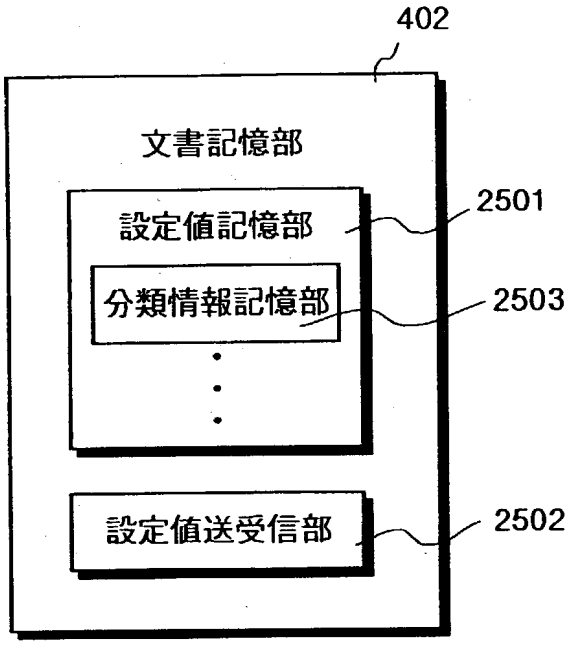
【図２３】


2301
番号
受付日
営業所
車種
年式
内容
クラスタ番号
2 1997/3/5 富山 ABC2000 1995 排気が黒い 1
11 1997/3/6 東京 ABC1600 1994 排気が黒い 1
15 1997/3/7 札幌 ABC2000 1996 排気が黒い 1
23 1997/3/7 福岡 ABC2000 1995 排気が黒い 1
35 1997/3/8 長野 K-M1200 1992 排気がくさい 1
54 1997/3/10 東京 ABC1600 1994 排気が黒い 1
63 1997/3/12 長野 ABC2000 1996 排気が黒い 1
73 1997/3/14 東京 DEF1600 1997 排気がにおう 1
82 1997/3/14 福岡 ABC1600 1992 排気が黒い 1
14 1997/3/6 仙台 XYZ3000 1995 暖房が効かない 1
18 1997/3/7 長野 XYZ3000 1997 暖房が効かず、寒い 2
22 1997/3/7 長野 XYZ3000 1997 暖房が効かない 2
27 1997/3/8 仙台 XYZ3000 1995 暖房の効きが悪い 2
37 1997/3/8 長野 XYZ3000 1995 暖房が効かない 2

【図２４】

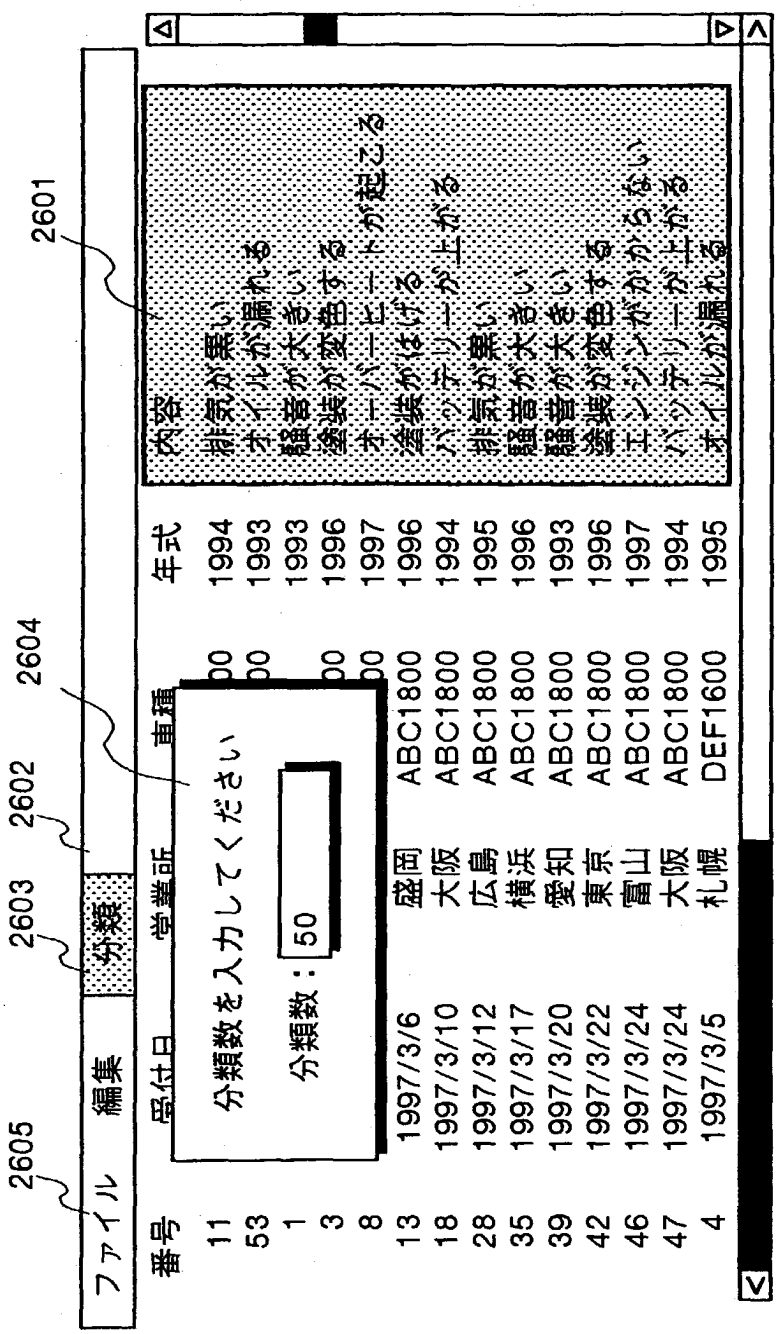

番号
受付日
営業所
車種
年式
内容
クラスタ番号
2401
件数
ABC2000 ABC1600 KLM1200 DEF1600 車種

【図２５】

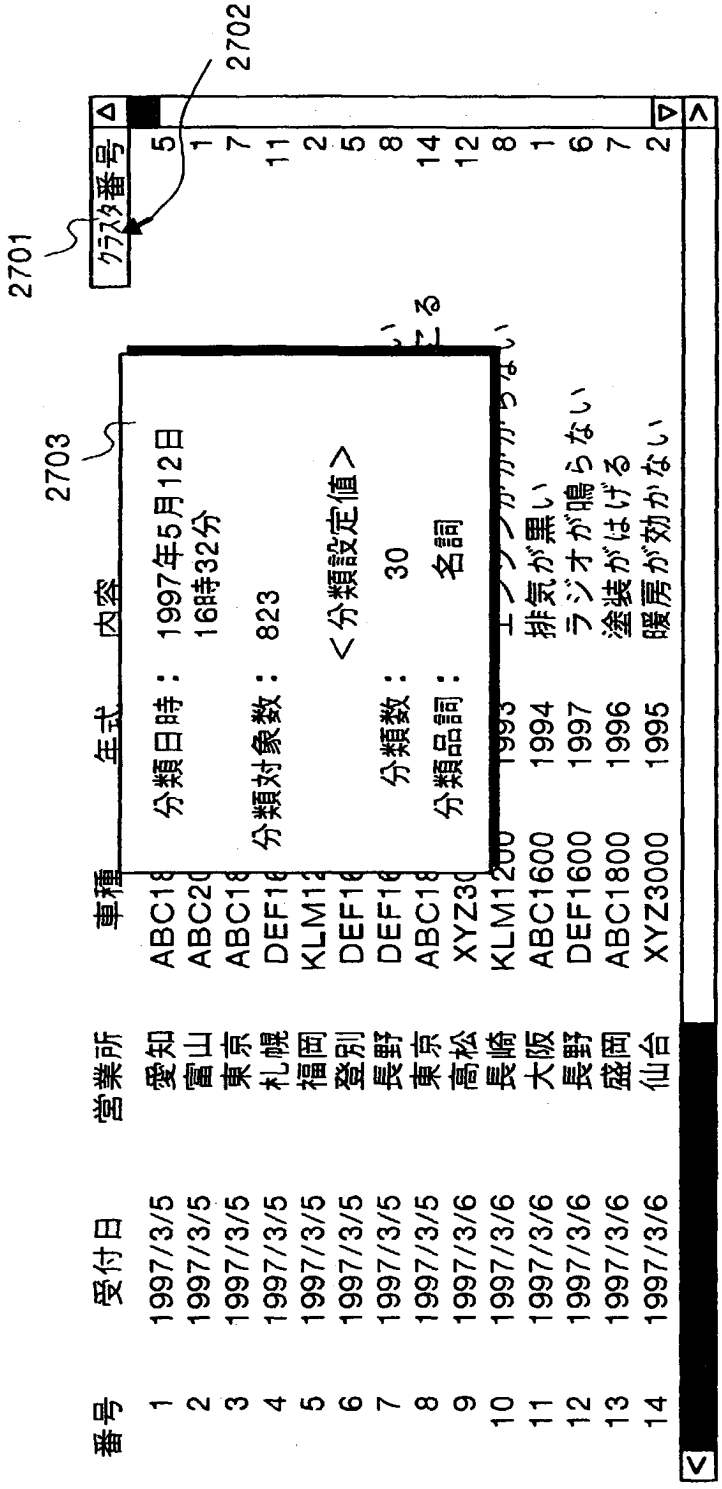


出証特平１１－３０９１７２６

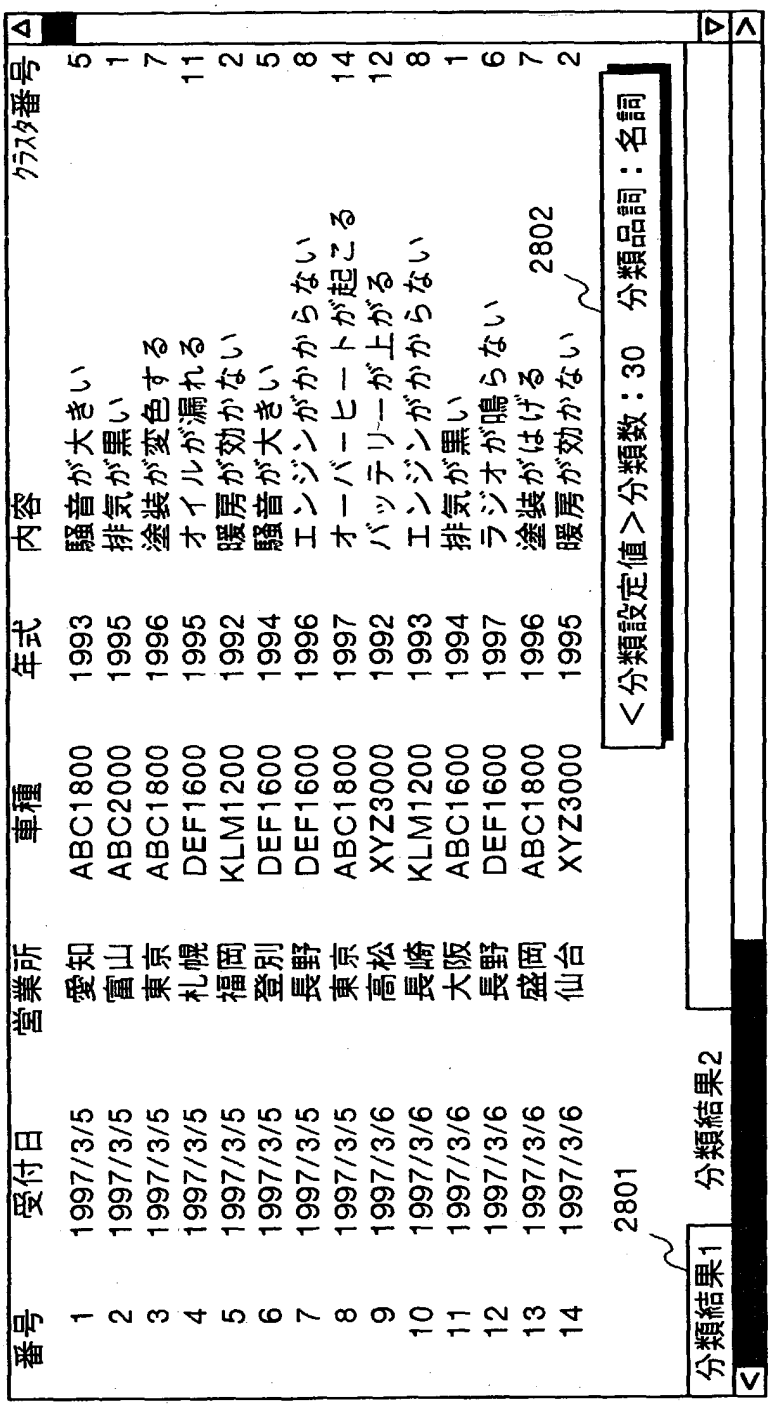
【図２６】


2601
2602
2603
2604
2605
ファイル
編集
分類
分類数を入力してください
分類数：
50
番号
受付日
営業部
直種
年式
内容
11 1997/3/6 盛岡 ABC1800 1994 排気が黒い
53 1997/3/10 大阪 ABC1800 1993 オイルが漏れる
1 1997/3/12 広島 ABC1800 1993 騒音が大きい
3 1997/3/17 横浜 ABC1800 1996 塗装が変色する
8 1997/3/20 愛知 ABC1800 1997 オーバーヒートが起こる
13 1997/3/22 東京 ABC1800 1996 塗装がはげる
18 1997/3/24 富山 ABC1800 1994 バッテリーが上がる
28 1997/3/24 大阪 ABC1800 1995 排気が黒い
35 1997/3/5 札幌 DEF1600 1996 騒音が大きい
39 1993 塗装が変色する
42 1996 オーバーヒートが起こる
46 1997 バッテリーが上がる
47 1994 オイルが漏れる
4 1995

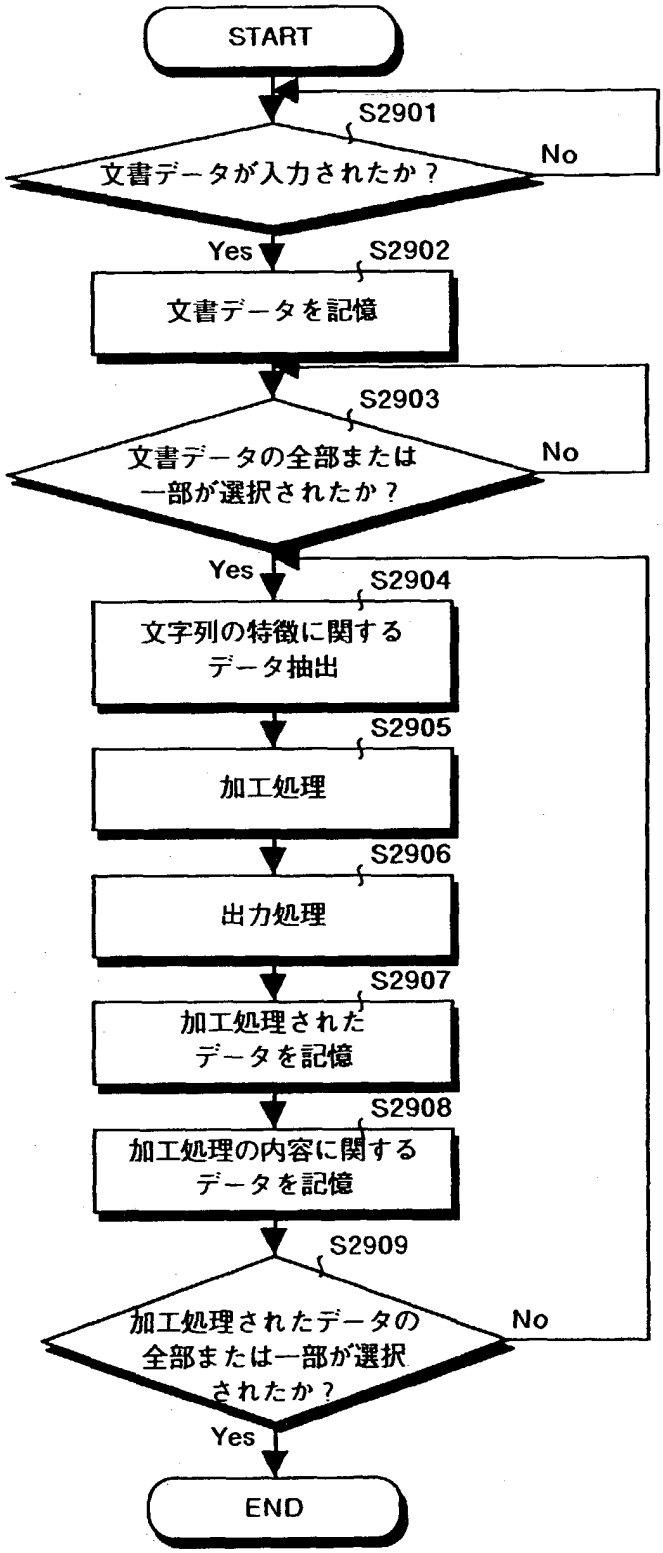
【図２７】


2701
2702
2703
クラスタ番号
番号
受付日
営業所
車種
年式
内容
分類日時：1997年5月12日
16時32分
分類対象数：823
＜分類設定値＞
分類数：30
分類品詞：名詞

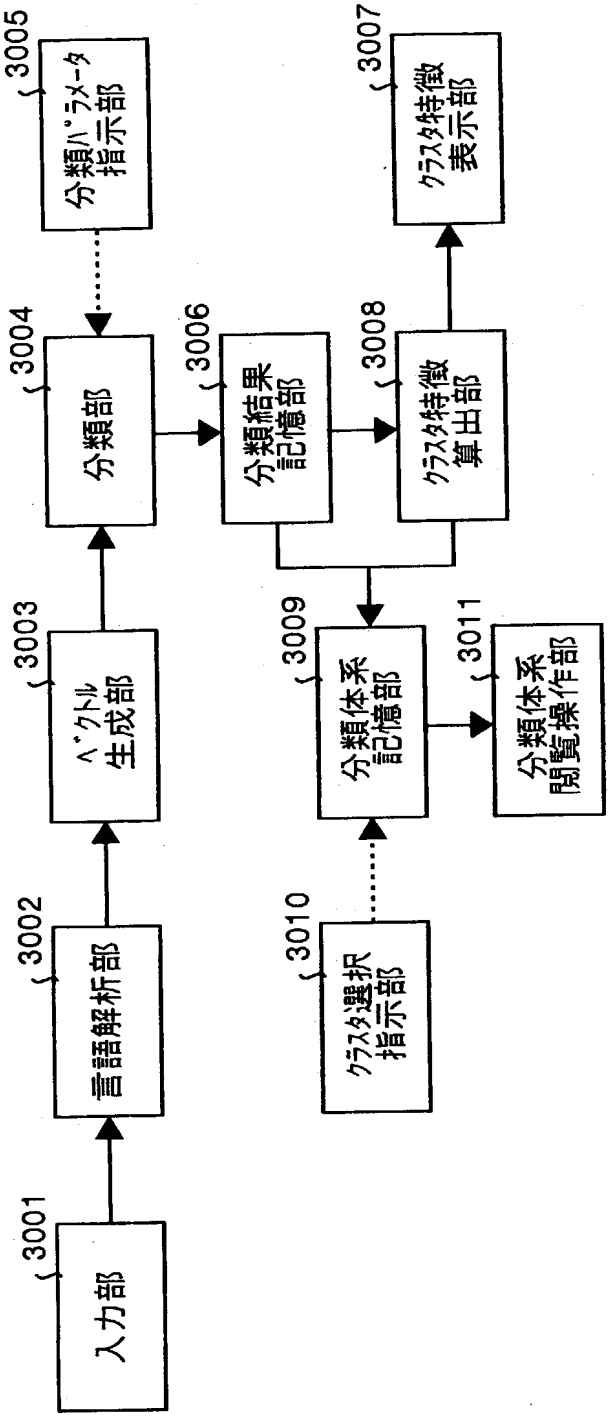
【図２８】


番号
受付日
営業所
車種
年式
内容
クラスタ番号
1 1997/3/5 愛知 ABC1800 1993 騒音が大きい 5
2 1997/3/5 富山 ABC2000 1995 排気が黒い 1
3 1997/3/5 東京 ABC1800 1996 塗装が変色する 7
4 1997/3/5 札幌 DEF1600 1995 オイルが漏れる 11
5 1997/3/5 福岡 KLM1200 1992 暖房が効かない 2
6 1997/3/5 登別 DEF1600 1994 騒音が大きい 5
7 1997/3/5 長野 DEF1600 1996 エンジンがかからない 8
8 1997/3/5 東京 ABC1800 1997 オーバーヒートが起こる 14
9 1997/3/6 高松 XYZ3000 1992 バッテリーが上がる 12
10 1997/3/6 長崎 KLM1200 1993 エンジンがかからない 8
11 1997/3/6 大阪 ABC1600 1994 排気が黒い 1
12 1997/3/6 長野 DEF1600 1997 ラジオが鳴らない 6
13 1997/3/6 盛岡 ABC1800 1996 塗装がはげる 7
14 1997/3/6 仙台 XYZ3000 1995 暖房が効かない 2
2801
2802
＜分類設定値＞分類数：30 分類品詞：名詞
分類結果1
分類結果2

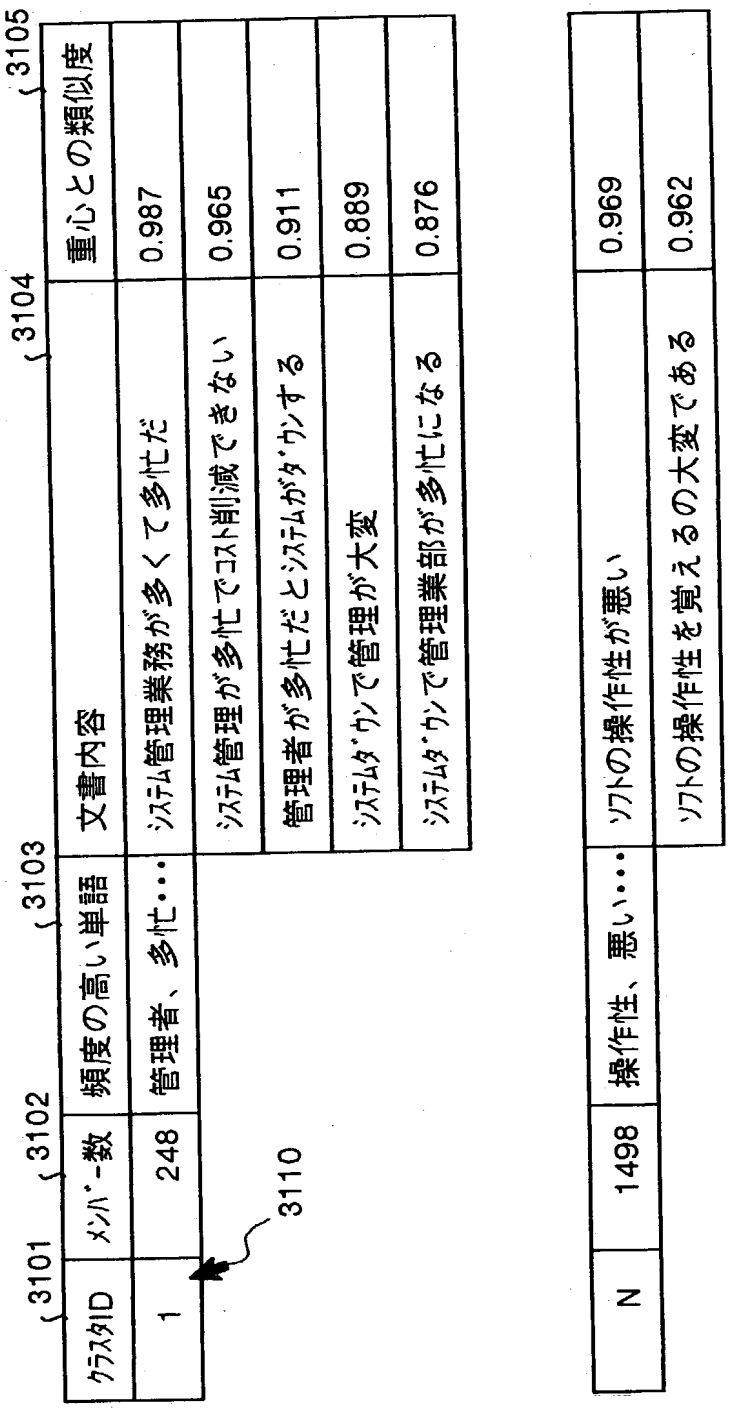
【図２９】


START
S2901
文書データが入力されたか？
No
Yes
S2902
文書データを記憶
S2903
文書データの全部または
一部が選択されたか？
No
Yes
S2904
文字列の特徴に関する
データ抽出
S2905
加工処理
S2906
出力処理
S2907
加工処理された
データを記憶
S2908
加工処理の内容に関する
データを記憶
S2909
加工処理されたデータの
全部または一部が選択
されたか？
No
Yes
END

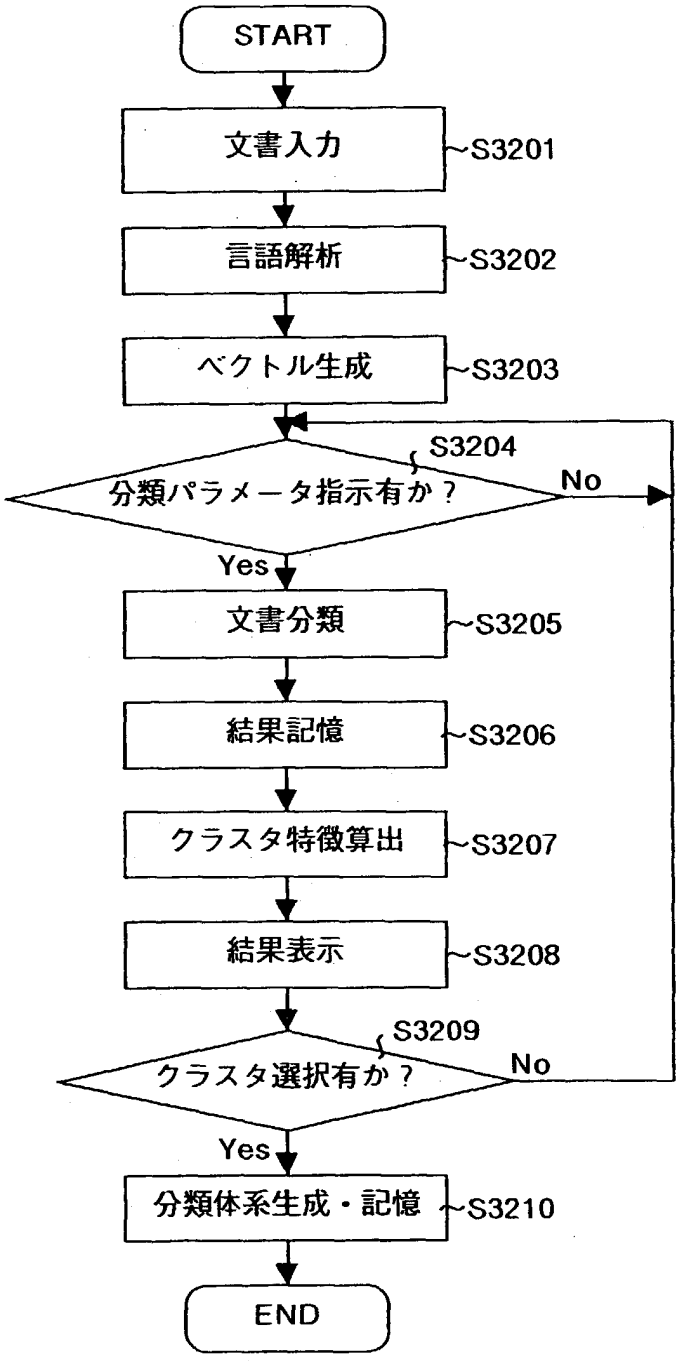
【図３０】


3001
入力部
3002
言語解析部
3003
ベクトル生成部
3004
分類部
3005
分類パラメータ指示部
3006
分類結果記憶部
3007
クラスタ特徴表示部
3008
クラスタ特徴算出部
3009
分類体系記憶部
3010
クラスタ選択指示部
3011
分類体系閲覧操作部

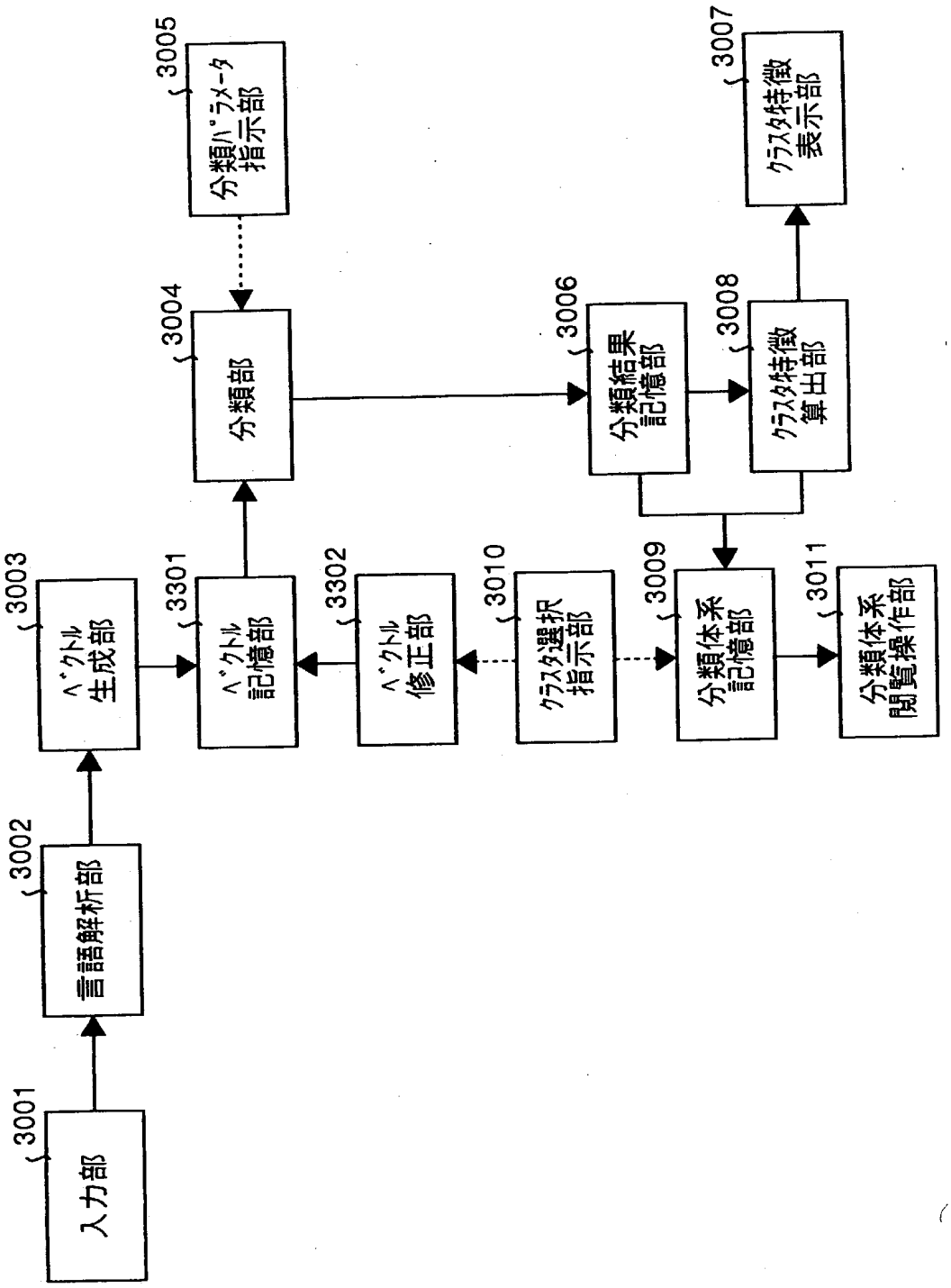
【図３１】


3101
3102
3103
3104
3105
クラスタID
メンバー数
頻度の高い単語
文書内容
重心との類似度
1
248
管理者、多忙・・・
システム管理業務が多くて多忙だ
0.987
システム管理が多忙でコスト削減できない
0.965
管理者が多忙だとシステムがダウンする
0.911
システムダウンで管理が大変
0.889
システムダウンで管理業部が多忙になる
0.876
3110
N
1498
操作性、悪い・・・
ソフトの操作性が悪い
0.969
ソフトの操作性を覚えるの大変である
0.962

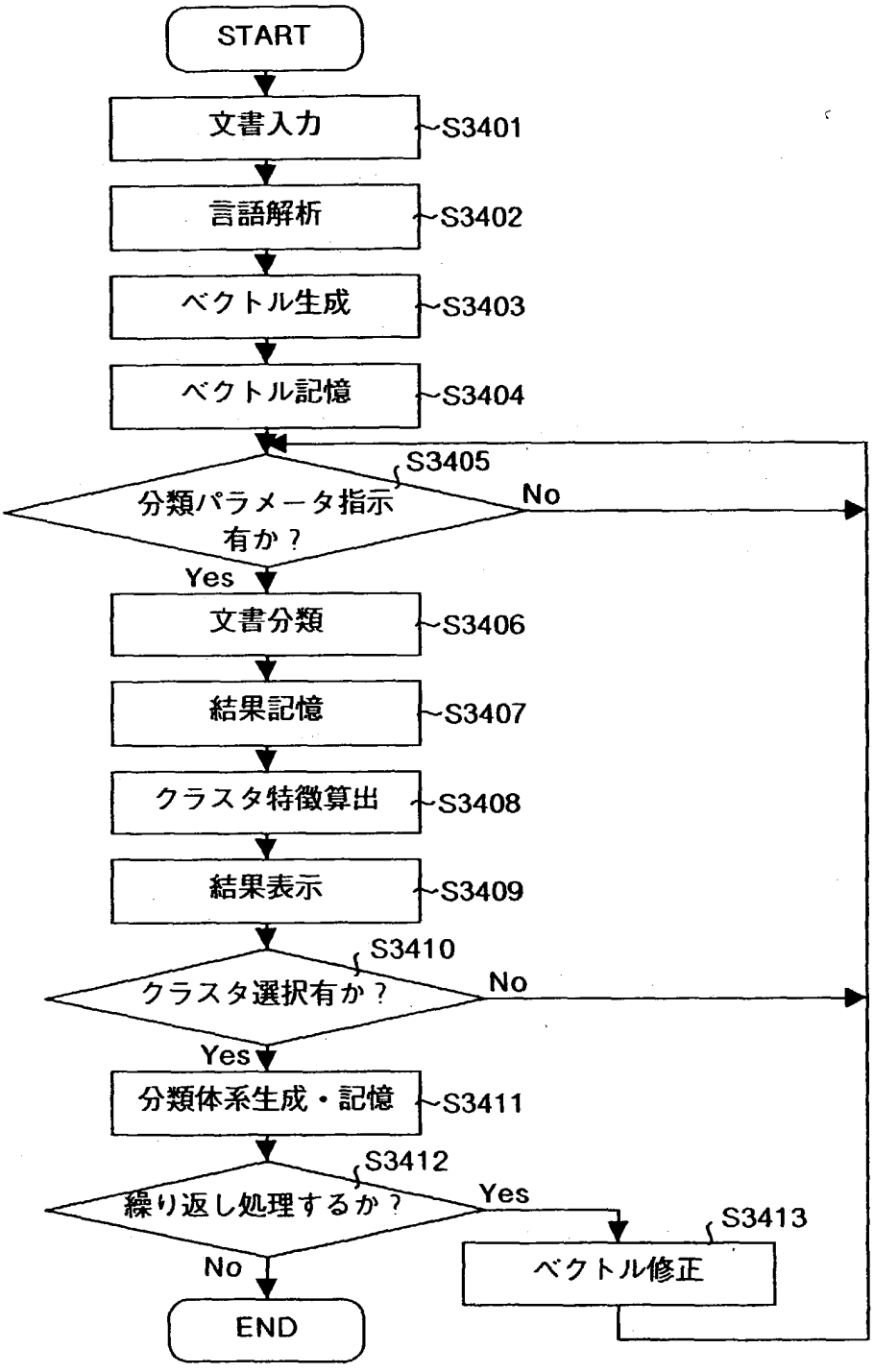
【図３２】


START
文書入力
S3201
言語解析
S3202
ベクトル生成
S3203
S3204
分類パラメータ指示有か？
No
Yes
文書分類
S3205
結果記憶
S3206
クラスタ特徴算出
S3207
結果表示
S3208
S3209
クラスタ選択有か？
No
Yes
分類体系生成・記憶
S3210
END

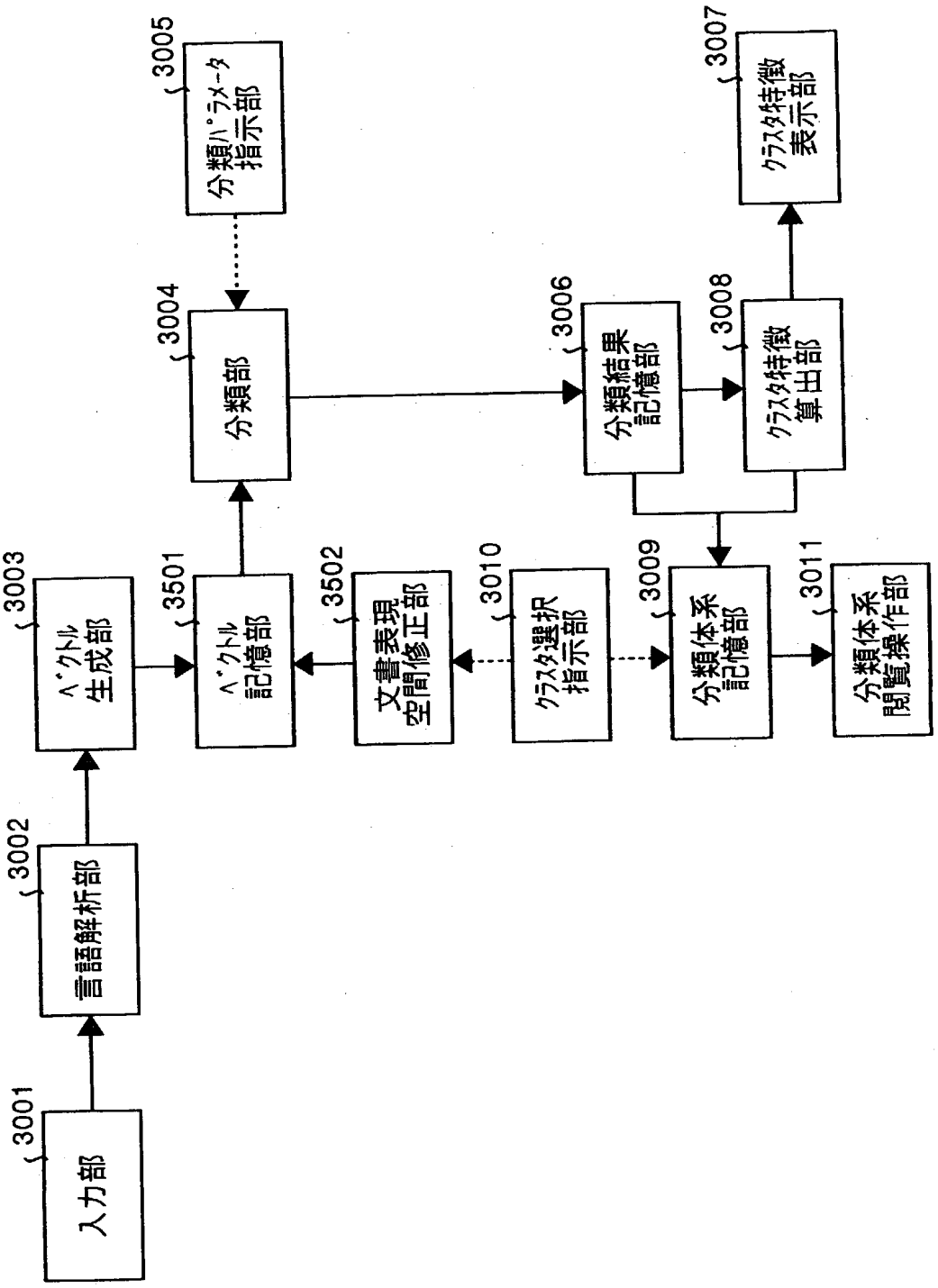
【図３３】


3001
入力部
3002
言語解析部
3003
ﾍﾞｸﾄﾙ生成部
3004
分類部
3005
分類ﾊﾟﾗﾒｰﾀ指示部
3006
分類結果記憶部
3007
クラスタ特徴表示部
3008
クラスタ特徴算出部
3009
分類体系記憶部
3010
クラスタ選択指示部
3011
分類体系閲覧操作部
3301
ﾍﾞｸﾄﾙ記憶部
3302
ﾍﾞｸﾄﾙ修正部

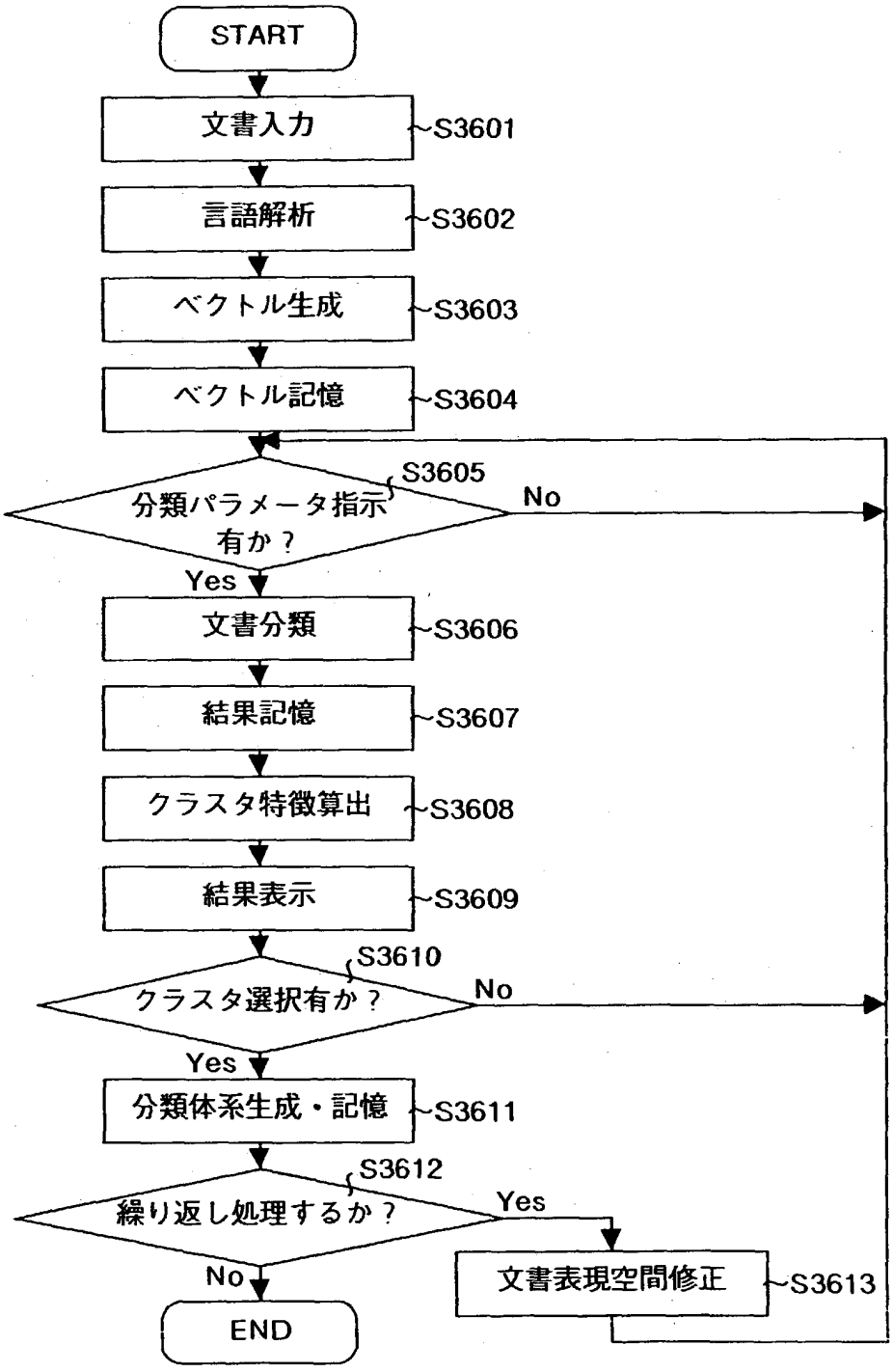
【図３４】


START
文書入力
S3401
言語解析
S3402
ベクトル生成
S3403
ベクトル記憶
S3404
S3405
分類パラメータ指示有か？
No
Yes
文書分類
S3406
結果記憶
S3407
クラスタ特徴算出
S3408
結果表示
S3409
S3410
クラスタ選択有か？
No
Yes
分類体系生成・記憶
S3411
S3412
繰り返し処理するか？
Yes
S3413
ベクトル修正
No
END

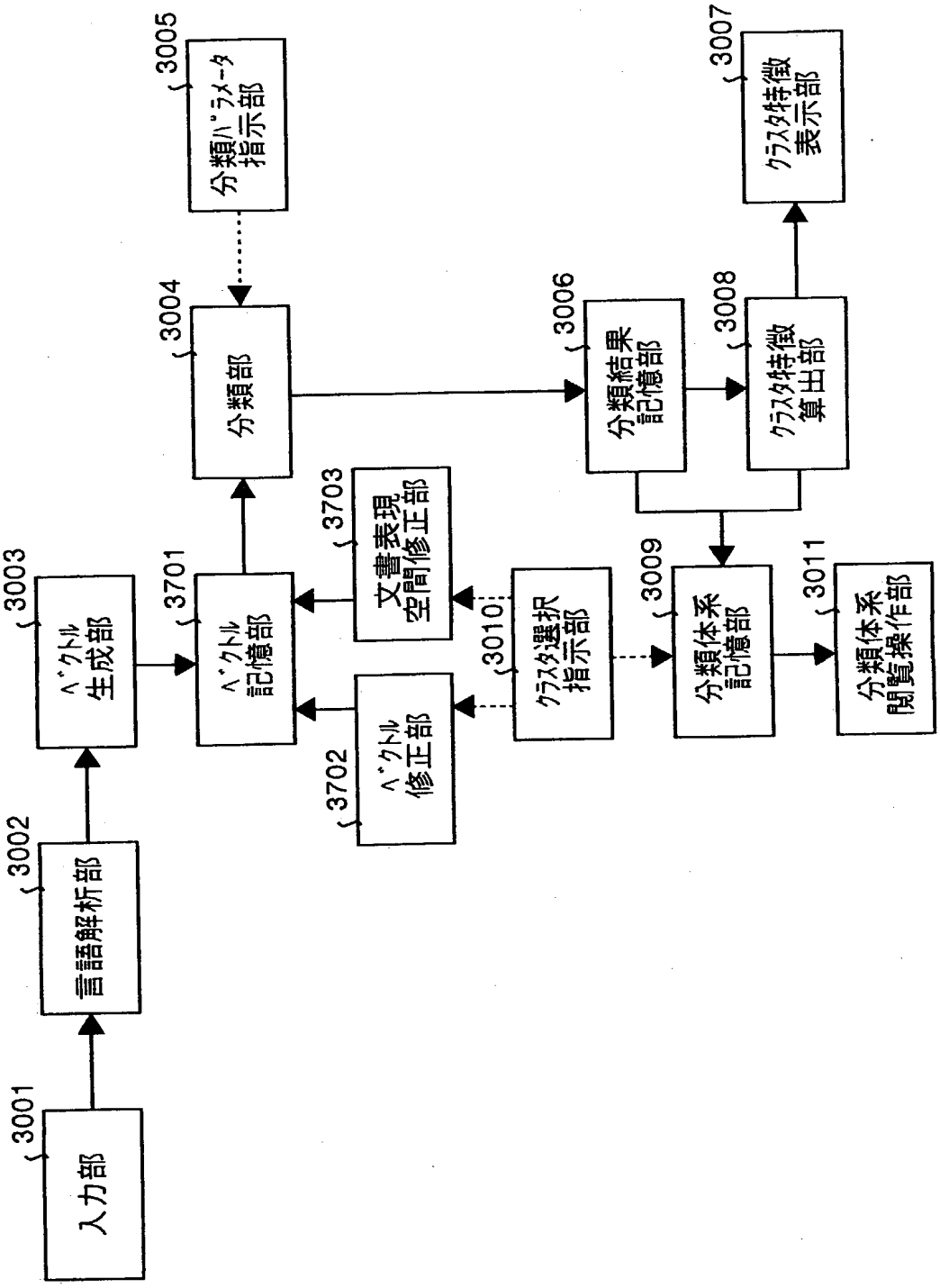
【図３５】


3001
入力部
3002
言語解析部
3003
ﾍﾞｸﾄﾙ
生成部
3004
分類部
3005
分類ﾊﾟﾗﾒｰﾀ
指示部
3006
分類結果
記憶部
3007
クラスタ特徴
表示部
3008
クラスタ特徴
算出部
3009
分類体系
記憶部
3010
クラスタ選択
指示部
3011
分類体系
閲覧操作部
3501
ﾍﾞｸﾄﾙ
記憶部
3502
文書表現
空間修正部

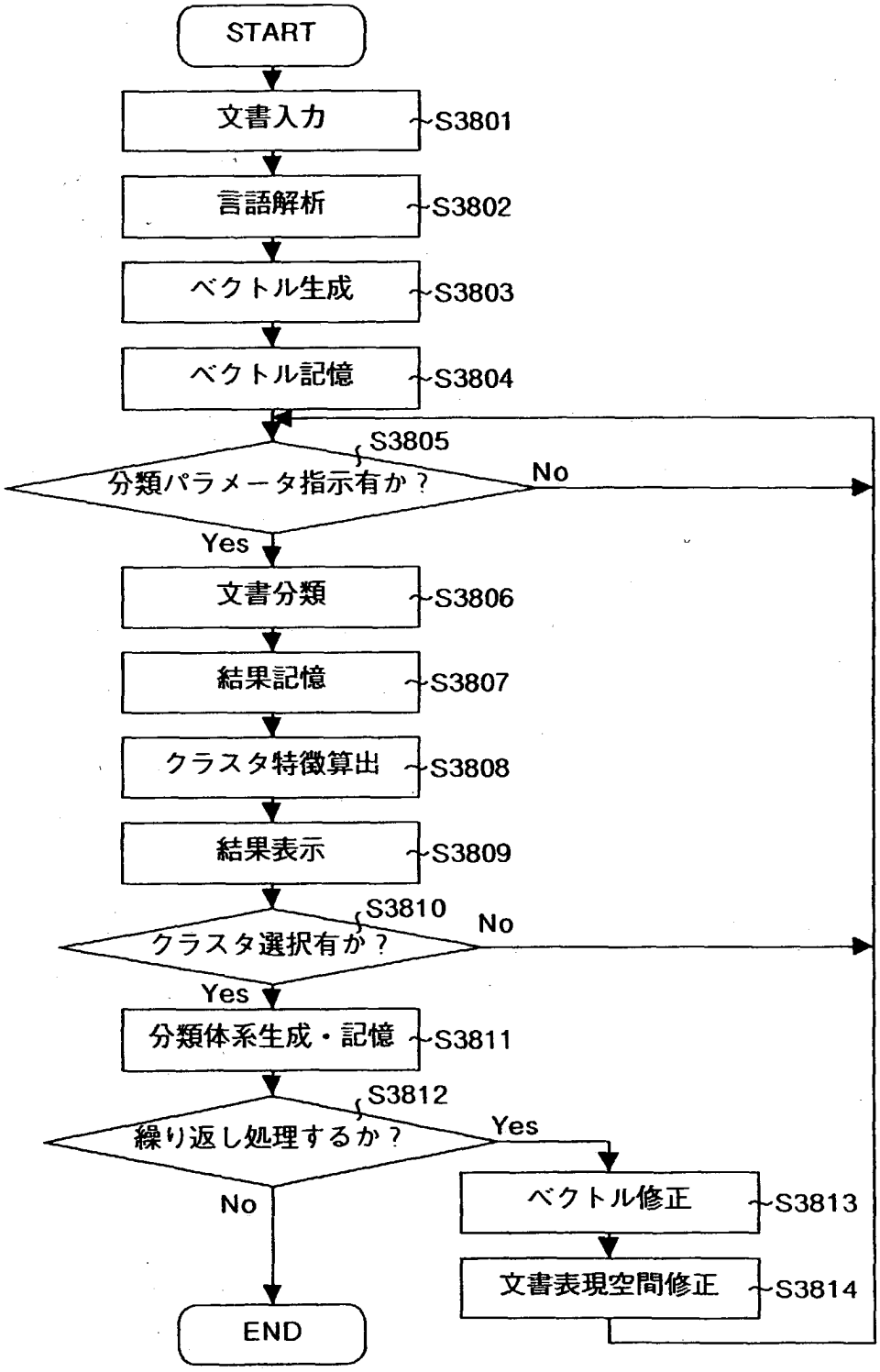
【図３６】


START
文書入力
S3601
言語解析
S3602
ベクトル生成
S3603
ベクトル記憶
S3604
S3605
分類パラメータ指示有か？
No
Yes
文書分類
S3606
結果記憶
S3607
クラスタ特徴算出
S3608
結果表示
S3609
S3610
クラスタ選択有か？
No
Yes
分類体系生成・記憶
S3611
S3612
繰り返し処理するか？
Yes
No
文書表現空間修正
S3613
END

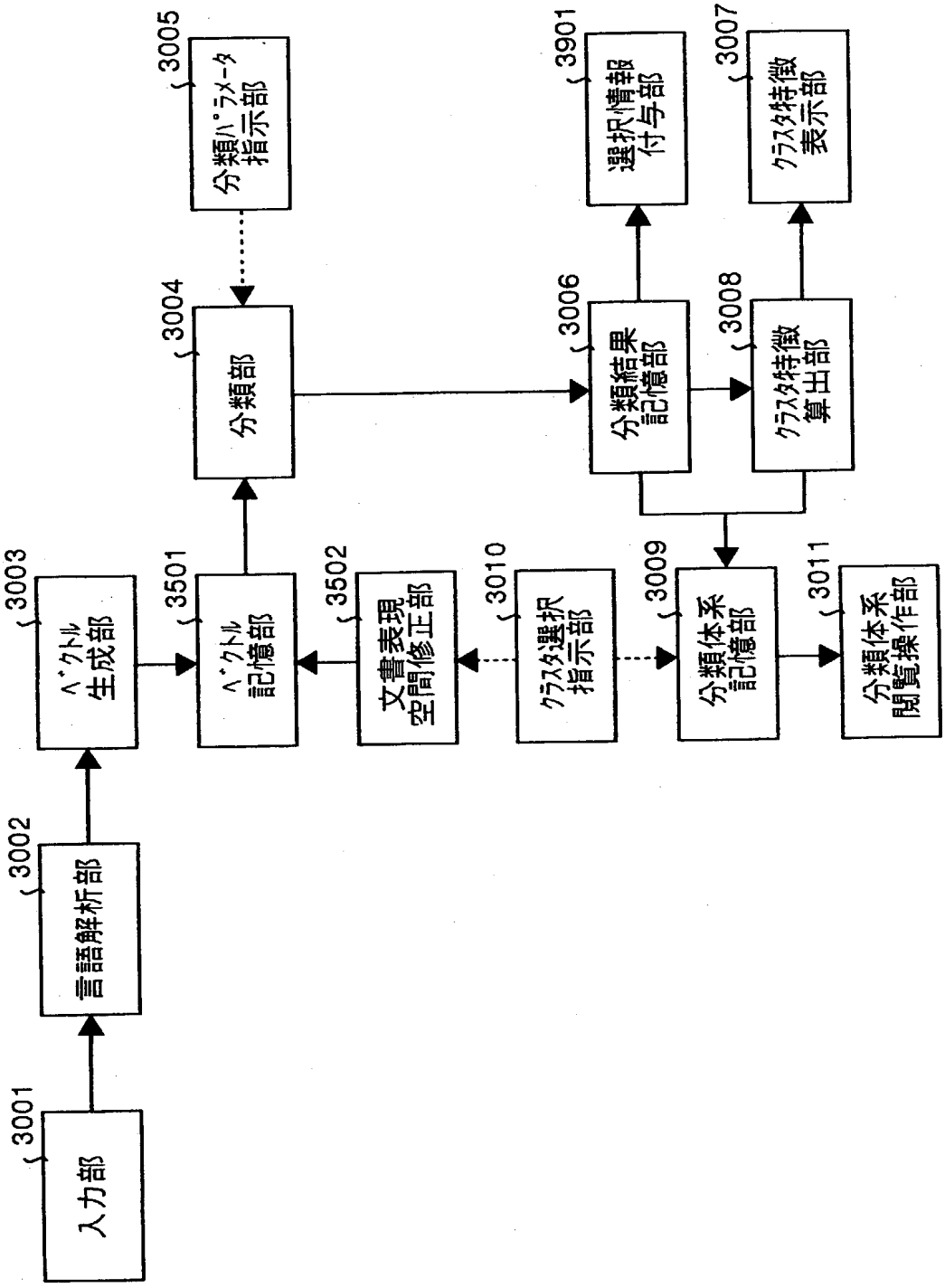
【図３７】


3001
入力部
3002
言語解析部
3003
ﾍﾞｸﾄﾙ
生成部
3004
分類部
3005
分類ﾊﾟﾗﾒｰﾀ
指示部
3006
分類結果
記憶部
3007
ｸﾗｽﾀ特徴
表示部
3008
ｸﾗｽﾀ特徴
算出部
3009
分類体系
記憶部
3010
ｸﾗｽﾀ選択
指示部
3011
分類体系
閲覧操作部
3701
ﾍﾞｸﾄﾙ
記憶部
3702
ﾍﾞｸﾄﾙ
修正部
3703
文書表現
空間修正部

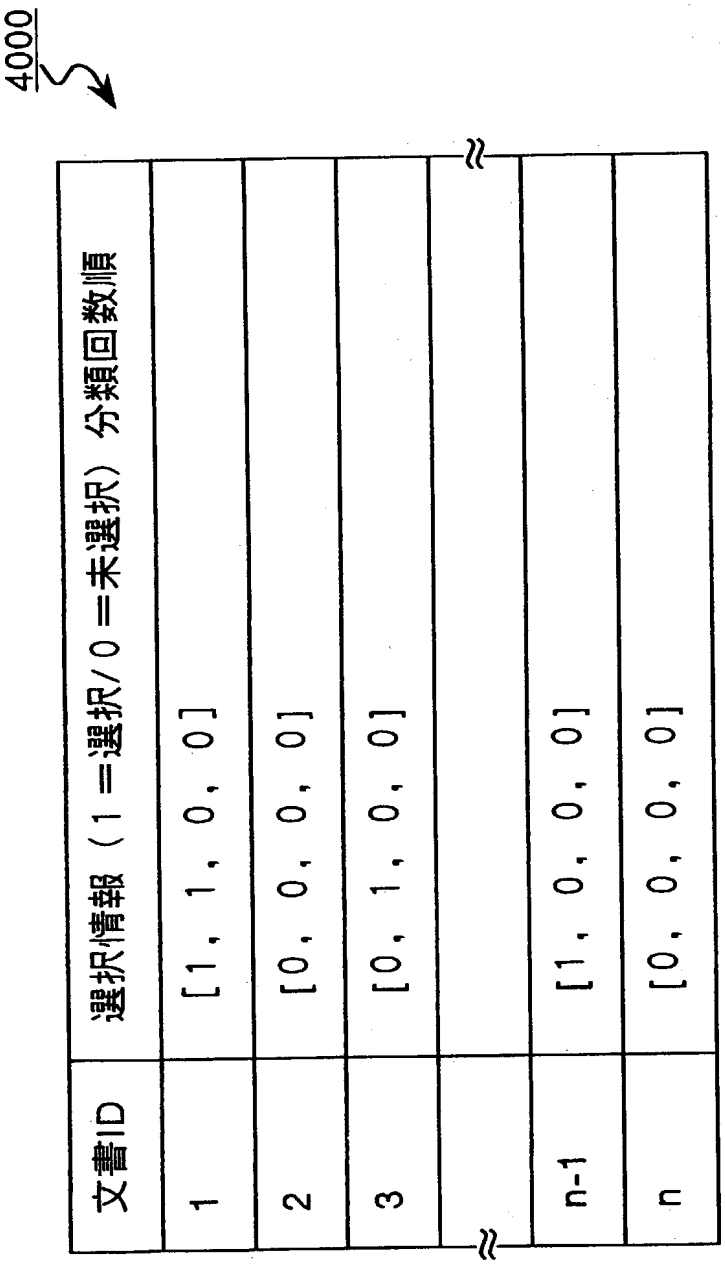
【図３８】


START
文書入力
S3801
言語解析
S3802
ベクトル生成
S3803
ベクトル記憶
S3804
S3805
分類パラメータ指示有か？
No
Yes
文書分類
S3806
結果記憶
S3807
クラスタ特徴算出
S3808
結果表示
S3809
S3810
クラスタ選択有か？
No
Yes
分類体系生成・記憶
S3811
S3812
繰り返し処理するか？
Yes
No
ベクトル修正
S3813
文書表現空間修正
S3814
END

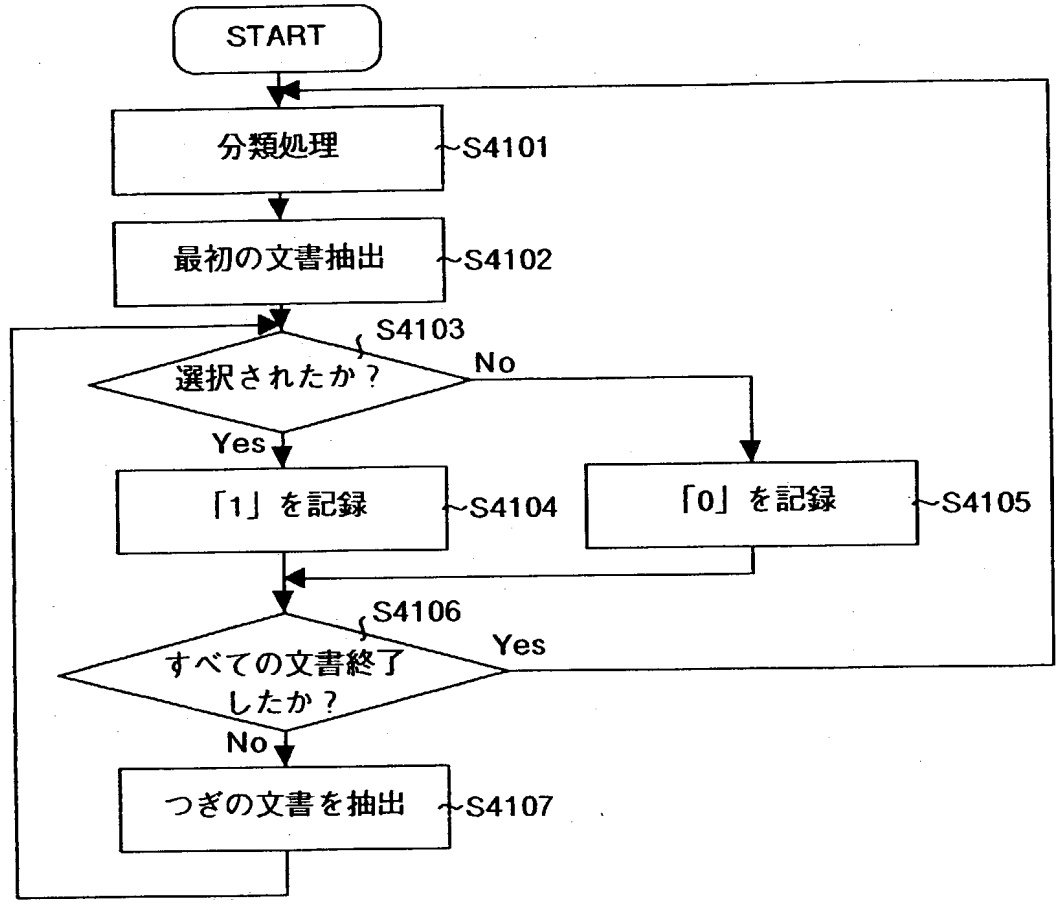
【図３９】


3001
入力部
3002
言語解析部
3003
ﾍﾞｸﾄﾙ
生成部
3004
分類部
3005
分類ﾊﾟﾗﾒｰﾀ
指示部
3501
ﾍﾞｸﾄﾙ
記憶部
3502
文書表現
空間修正部
3010
ｸﾗｽﾀ選択
指示部
3009
分類体系
記憶部
3011
分類体系
閲覧操作部
3006
分類結果
記憶部
3008
ｸﾗｽﾀ特徴
算出部
3901
選択情報
付与部
3007
ｸﾗｽﾀ特徴
表示部

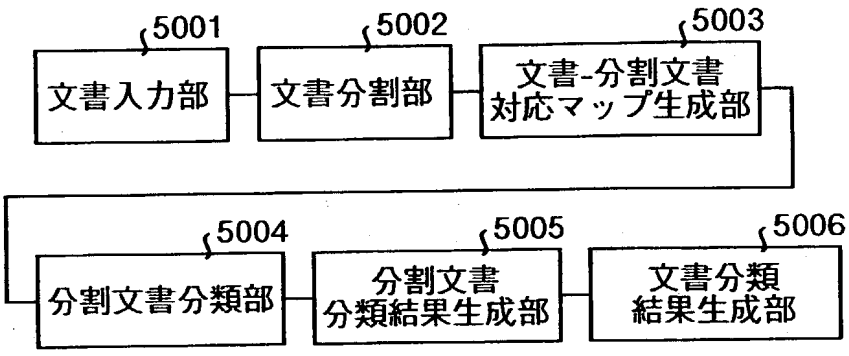
【図４０】


4000
文書ID
選択情報（１＝選択／０＝未選択）分類回数順
1
[1, 1, 0, 0]
2
[0, 0, 0, 0]
3
[0, 1, 0, 0]
n-1
[1, 0, 0, 0]
n
[0, 0, 0, 0]

【図４１】

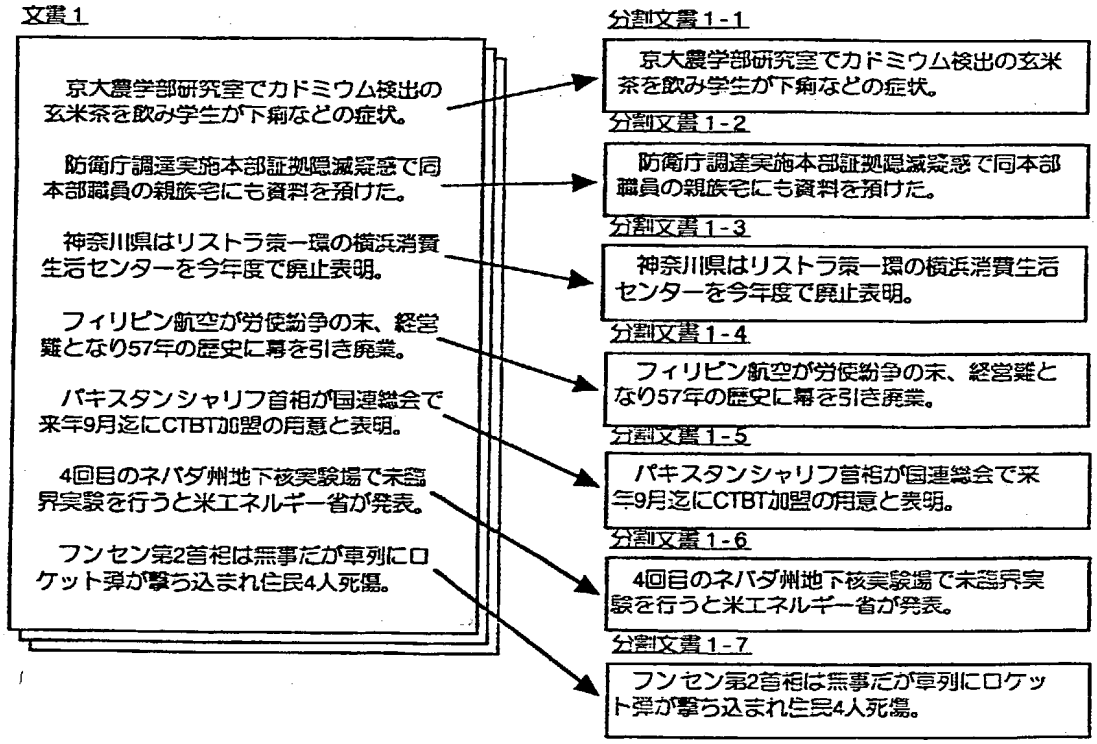

START
分類処理
S4101
最初の文書抽出
S4102
S4103
選択されたか？
No
Yes
「1」を記録
S4104
「0」を記録
S4105
S4106
すべての文書終了したか？
Yes
No
つぎの文書を抽出
S4107

【図４２】

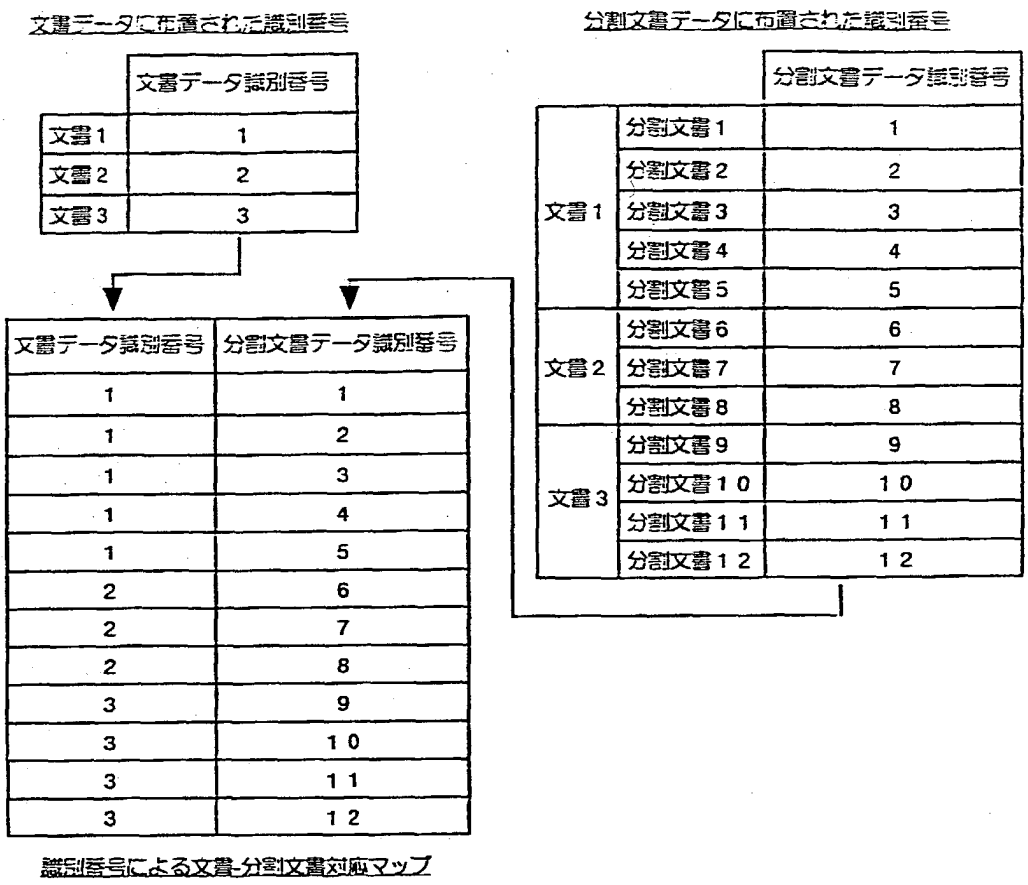


出証特平１１－３０９１７２６

【図４３】


文書１
京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。
分割文書１-１
京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
分割文書１-２
防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
分割文書１-３
神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
分割文書１-４
フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
分割文書１-５
パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
分割文書１-６
4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
分割文書１-７
フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。

【図４４】


文書データに布置された識別番号
文書データ識別番号
文書1 1
文書2 2
文書3 3
分割文書データに布置された識別番号
分割文書データ識別番号
文書1 分割文書1 1
分割文書2 2
分割文書3 3
分割文書4 4
分割文書5 5
文書2 分割文書6 6
分割文書7 7
分割文書8 8
文書3 分割文書9 9
分割文書10 10
分割文書11 11
分割文書12 12
文書データ識別番号 分割文書データ識別番号
1 1
1 2
1 3
1 4
1 5
2 6
2 7
2 8
3 9
3 10
3 11
3 12
識別番号による文書-分割文書対応マップ

【図４５】

分割文書データの特徴ベクトル表現

	分割文書データ識別番号	分割文書データ特徴ベクトル
分割文書１	１	（１、１、１）
分割文書２	２	（５、５、５）
分割文書３	３	（３、２、４）
分割文書４	４	（３、２、３）
分割文書５	５	（５、４、６）
分割文書６	６	（１、２、１）
分割文書７	７	（１、０、１）
分割文書８	８	（５、４、５）
分割文書９	９	（２、２、４）
分割文書１０	１０	（２、１、１）
分割文書１１	１１	（４、４、６）
分割文書１２	１２	（５、５、６）

文書分類（クラスタ分析手法を適用）

分割文書データを３つのカテゴリに分類した結果

	分割文書データ識別番号	分類カテゴリ	所属カテゴリの代表値との距離
分割文書１	１	カテゴリ１	０.２５
分割文書２	２	カテゴリ３	０.８７
分割文書３	３	カテゴリ２	０.４８
分割文書４	４	カテゴリ２	０.７４
分割文書５	５	カテゴリ３	０.５４
分割文書６	６	カテゴリ１	１.０３
分割文書７	７	カテゴリ１	１.０３
分割文書８	８	カテゴリ３	０.７０
分割文書９	９	カテゴリ２	０.７４
分割文書１０	１０	カテゴリ１	０.７５
分割文書１１	１１	カテゴリ３	０.９４
分割文書１２	１２	カテゴリ３	０.８３

分類カテゴリに関する情報

	代表値（所属分割文書データの重心）	所属データ数
カテゴリ１	（１.２５、１.００、１.００）	４
カテゴリ２	（２.６６、２.００、３.６６）	３
カテゴリ３	（４.８０、４.４０、５.６０）	５

分類カテゴリ間の距離

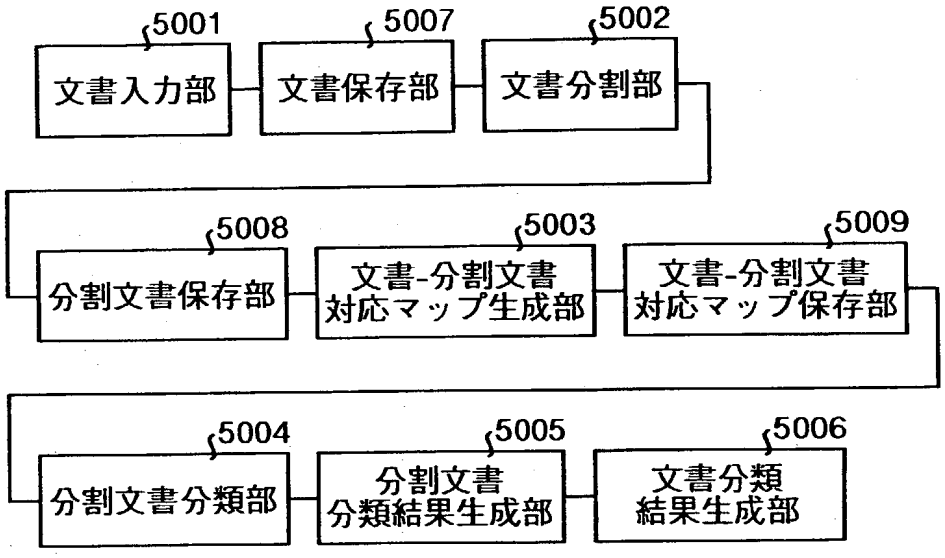
	カテゴリ２	カテゴリ３
カテゴリ１	３.１７	６.６８
カテゴリ２		３.６９

【図４６】

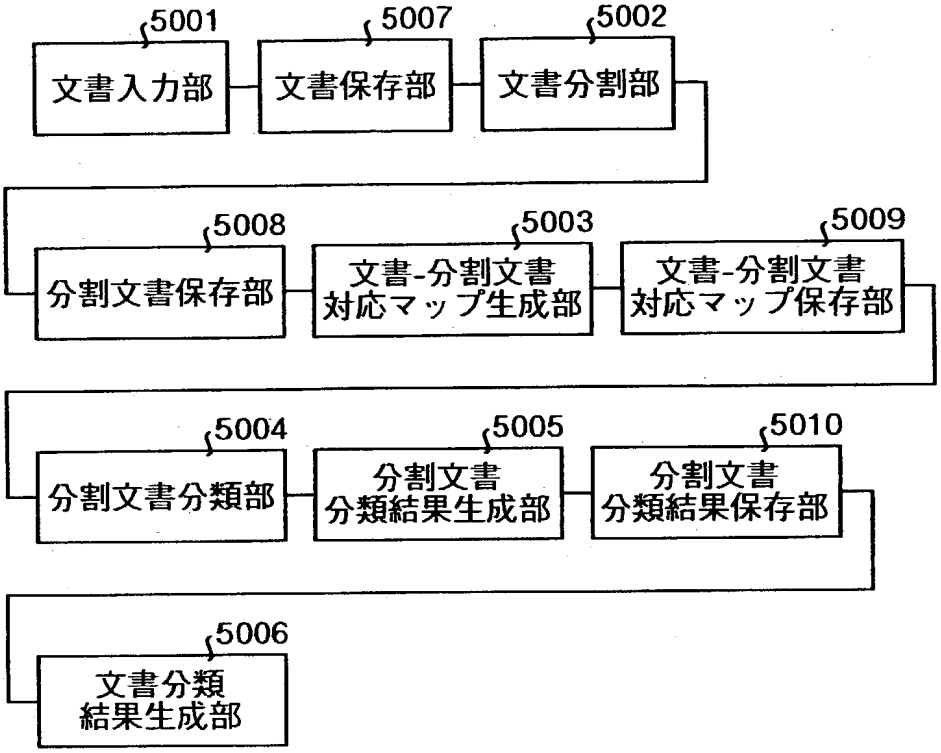
文書分類結果

分類カテゴリ	分割文書	類似度	所属文書	文書占有率	相対位置	類似順位
カテゴリ１	分割文書1	0.25	文書１	１／５	１／５	1
カテゴリ１	分割文書６	1.03	文書２	２／３	１／３	3
カテゴリ１	分割文書７	1.03	文書２	２／３	２／３	3
カテゴリ１	分割文書１０	0.75	文書３	１／４	２／４	2
カテゴリ２	分割文書３	0.48	文書１	２／５	３／５	1
カテゴリ２	分割文書４	0.74	文書１	２／５	４／５	2
カテゴリ２	分割文書９	0.74	文書３	１／４	１／４	2
カテゴリ３	分割文書２	0.87	文書１	２／５	２／５	4
カテゴリ３	分割文書５	0.54	文書１	２／５	５／５	1
カテゴリ３	分割文書８	0.70	文書２	１／３	３／３	2
カテゴリ３	分割文書１１	0.94	文書３	２／４	３／４	4
カテゴリ３	分割文書１２	0.83	文書３	２／４	４／４	3

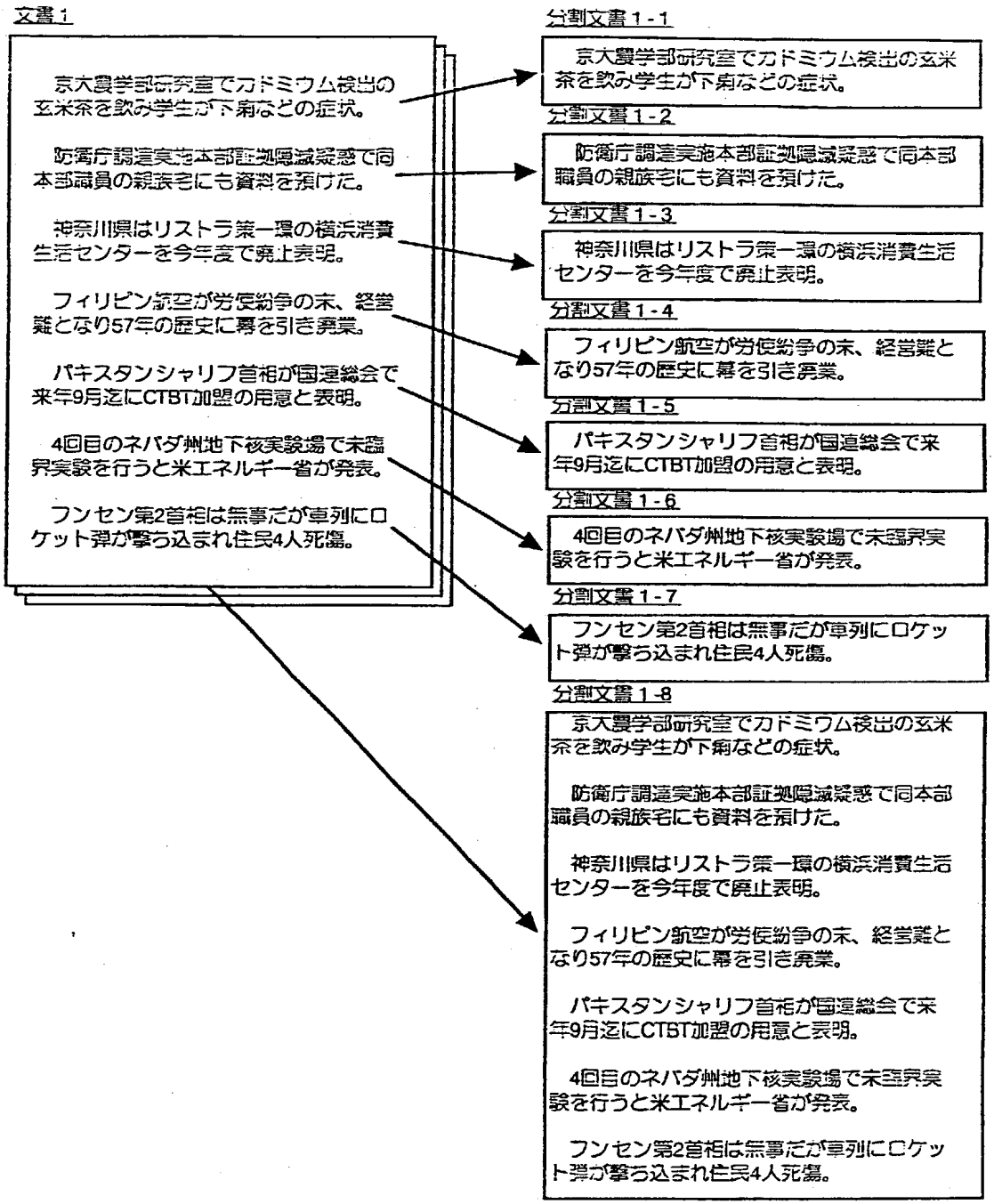
【図４７】


5001
文書入力部
5007
文書保存部
5002
文書分割部
5008
分割文書保存部
5003
文書-分割文書
対応マップ生成部
5009
文書-分割文書
対応マップ保存部
5004
分割文書分類部
5005
分割文書
分類結果生成部
5006
文書分類
結果生成部

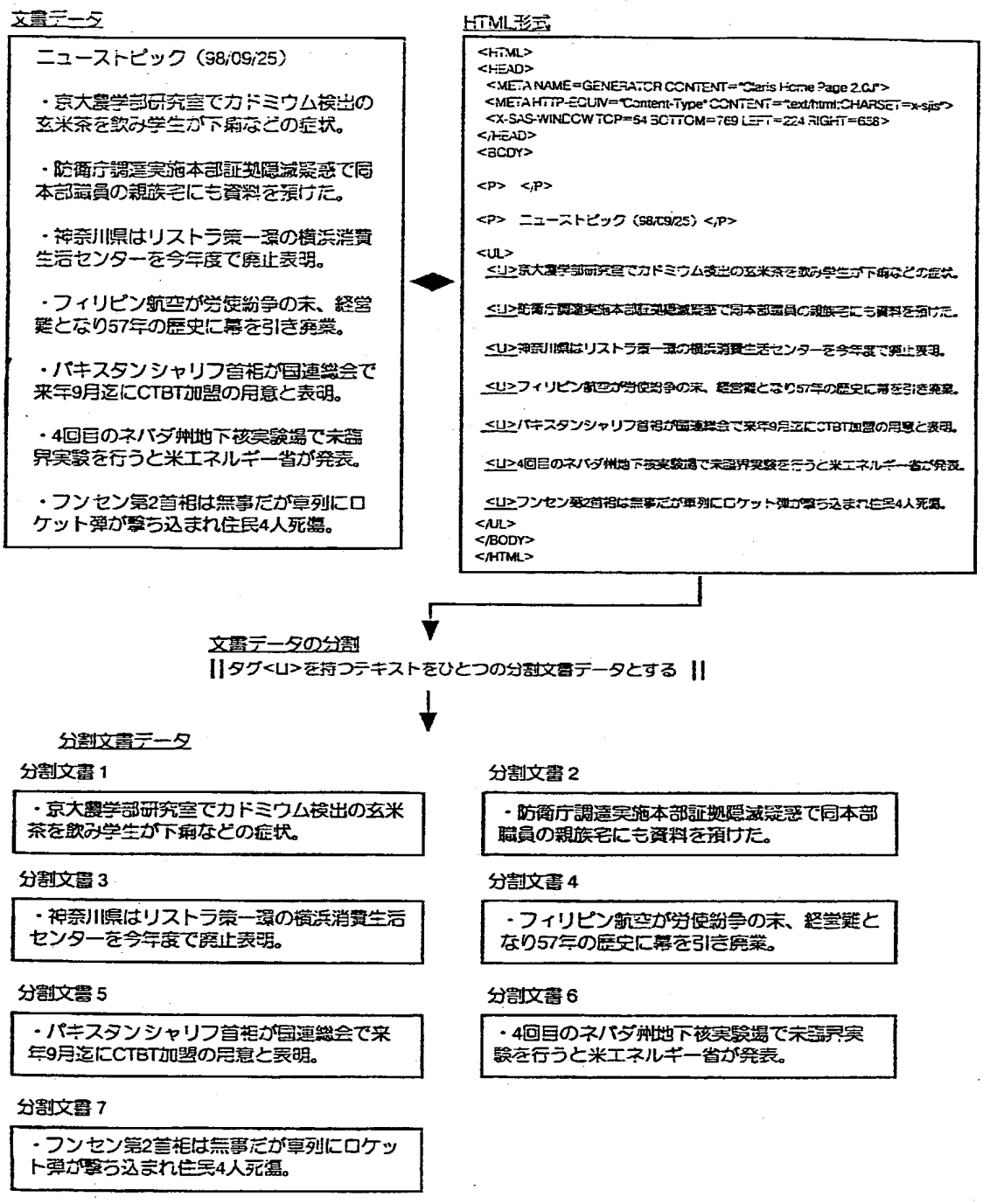
【図４８】


5001
文書入力部
5007
文書保存部
5002
文書分割部
5008
分割文書保存部
5003
文書-分割文書
対応マップ生成部
5009
文書-分割文書
対応マップ保存部
5004
分割文書分類部
5005
分割文書
分類結果生成部
5010
分割文書
分類結果保存部
5006
文書分類
結果生成部

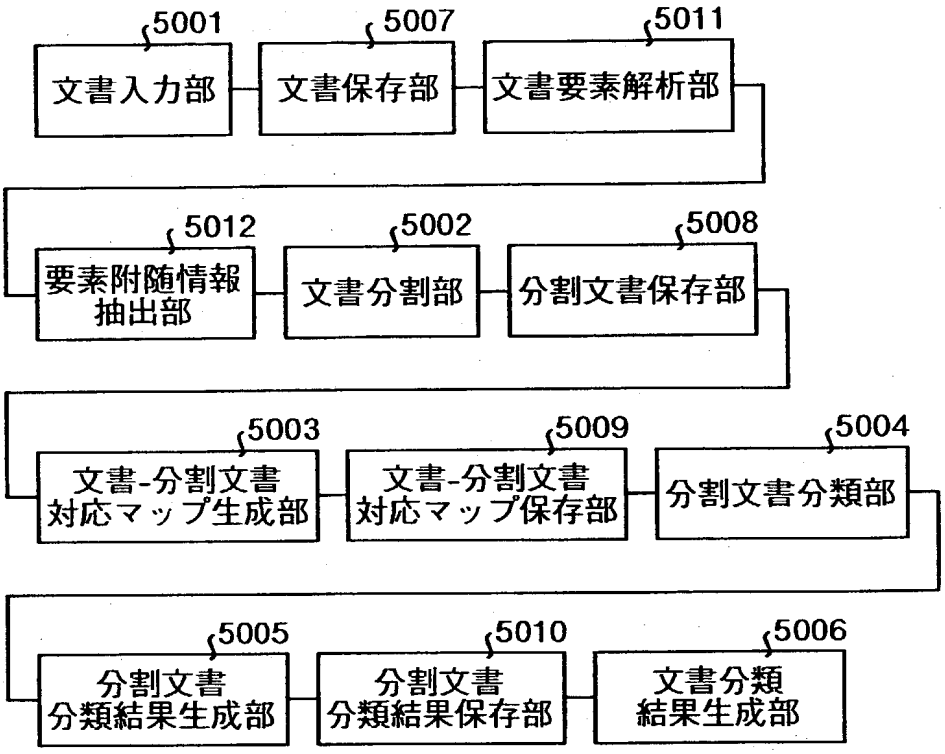
【図４９】


文書１
京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。
分割文書１-１
京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
分割文書１-２
防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
分割文書１-３
神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
分割文書１-４
フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
分割文書１-５
パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
分割文書１-６
4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
分割文書１-７
フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。
分割文書１-８
京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。

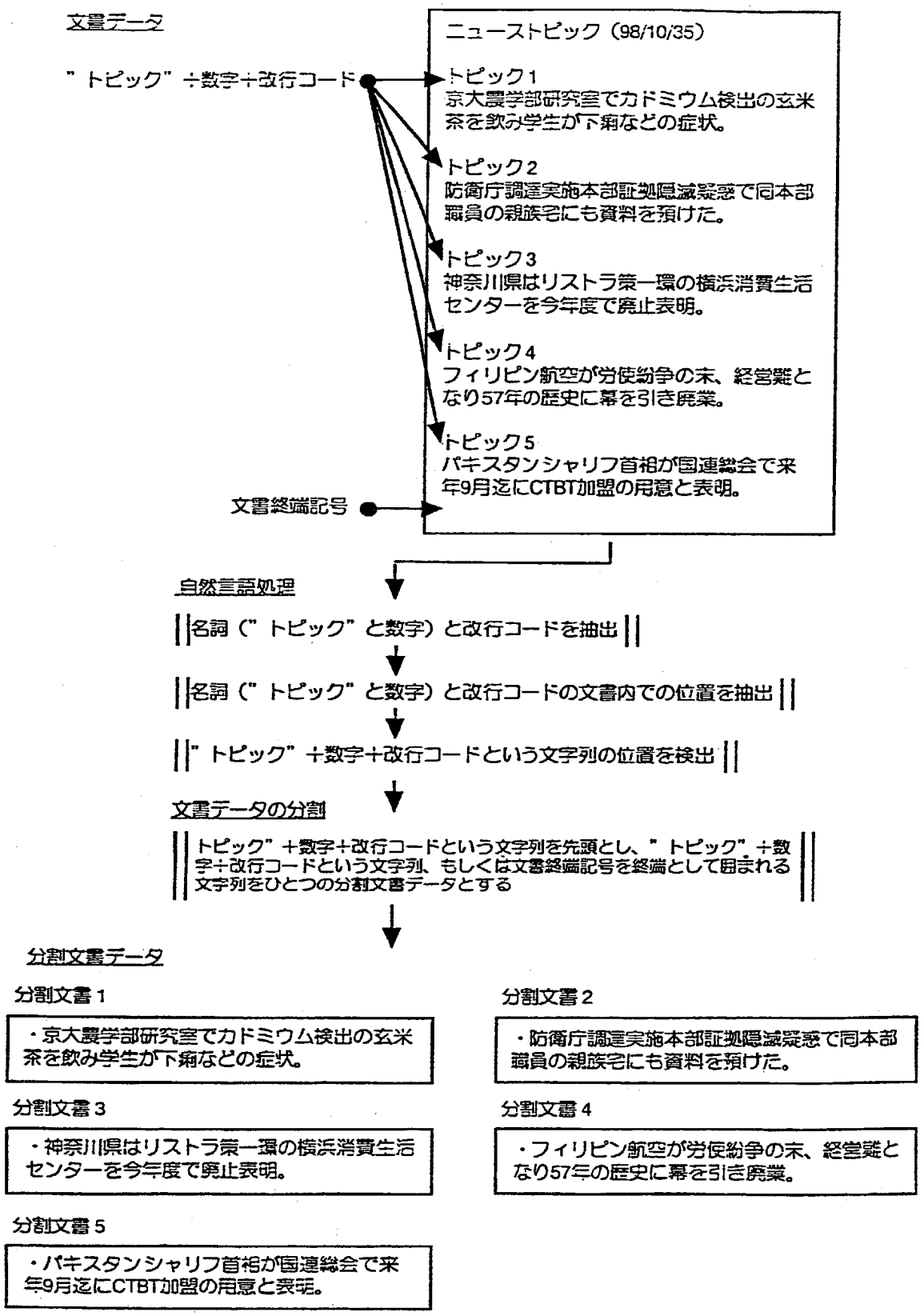
【図５０】


文書データ
ニューストピック（98/09/25）
・京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
・防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
・神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
・フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
・パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の同意と表明。
・4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
・フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。
HTML形式
<HTML>
<HEAD>
<META NAME=GENERATOR CONTENT="Claris Home Page 2.0J">
<META HTTP-EQUIV="Content-Type" CONTENT="text/html;CHARSET=x-sjis">
<X-SAS-WINDOW TOP=54 BOTTOM=769 LEFT=224 RIGHT=658>
</HEAD>
<BODY>
<P> </P>
<P> ニューストピック（98/09/25）</P>
<UL>
<LI>京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
<LI>防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
<LI>神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
<LI>フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
<LI>パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の同意と表明。
<LI>4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
<LI>フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。
</UL>
</BODY>
</HTML>
文書データの分割
||タグ<LI>を持つテキストをひとつの分割文書データとする ||
分割文書データ
分割文書１
・京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
分割文書２
・防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
分割文書３
・神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
分割文書４
・フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
分割文書５
・パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の同意と表明。
分割文書６
・4回目のネバダ州地下核実験場で未臨界実験を行うと米エネルギー省が発表。
分割文書７
・フンセン第2首相は無事だが車列にロケット弾が撃ち込まれ住民4人死傷。

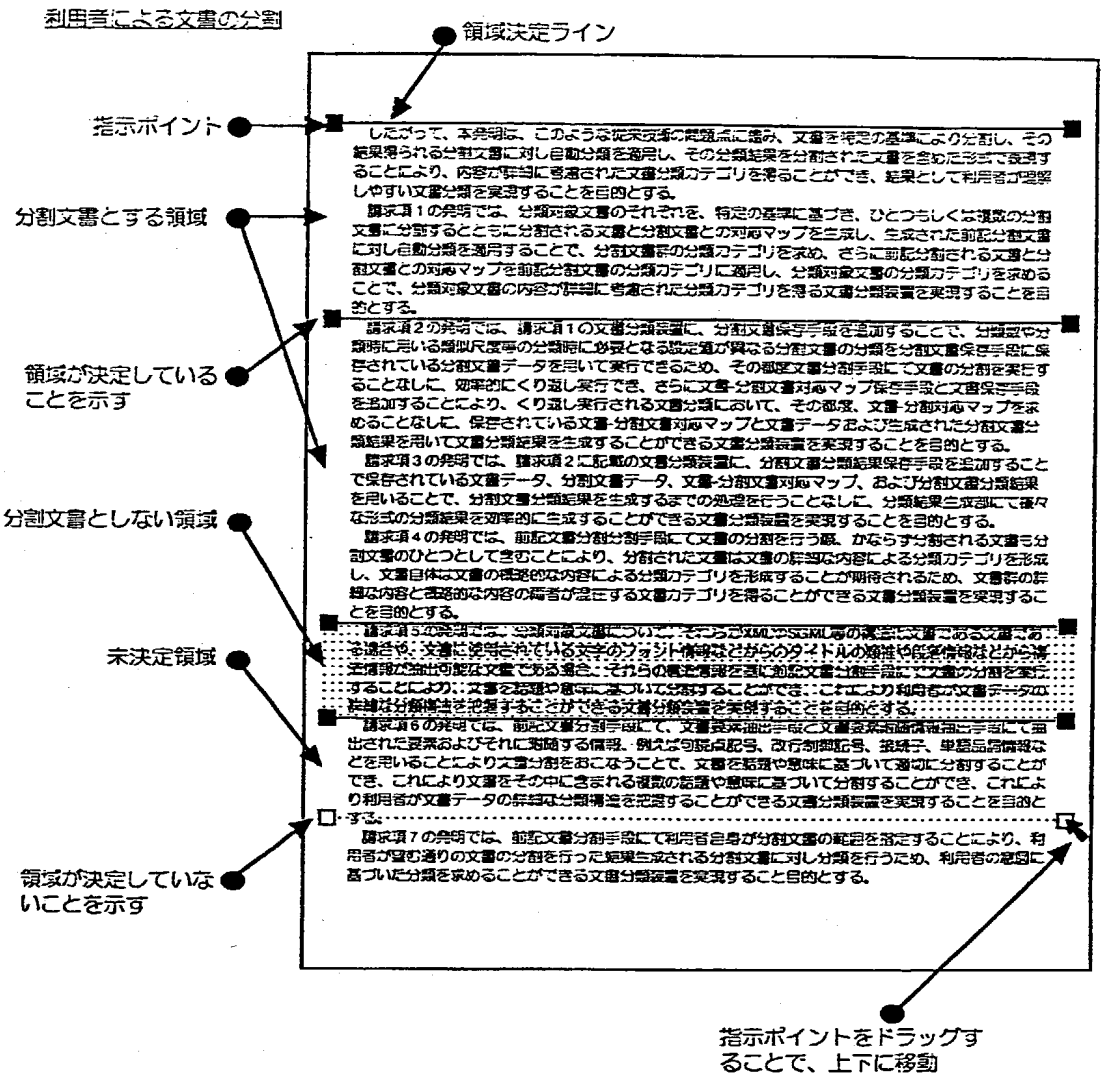
【図５１】


5001
文書入力部
5007
文書保存部
5011
文書要素解析部
5012
要素附随情報抽出部
5002
文書分割部
5008
分割文書保存部
5003
文書-分割文書対応マップ生成部
5009
文書-分割文書対応マップ保存部
5004
分割文書分類部
5005
分割文書分類結果生成部
5010
分割文書分類結果保存部
5006
文書分類結果生成部

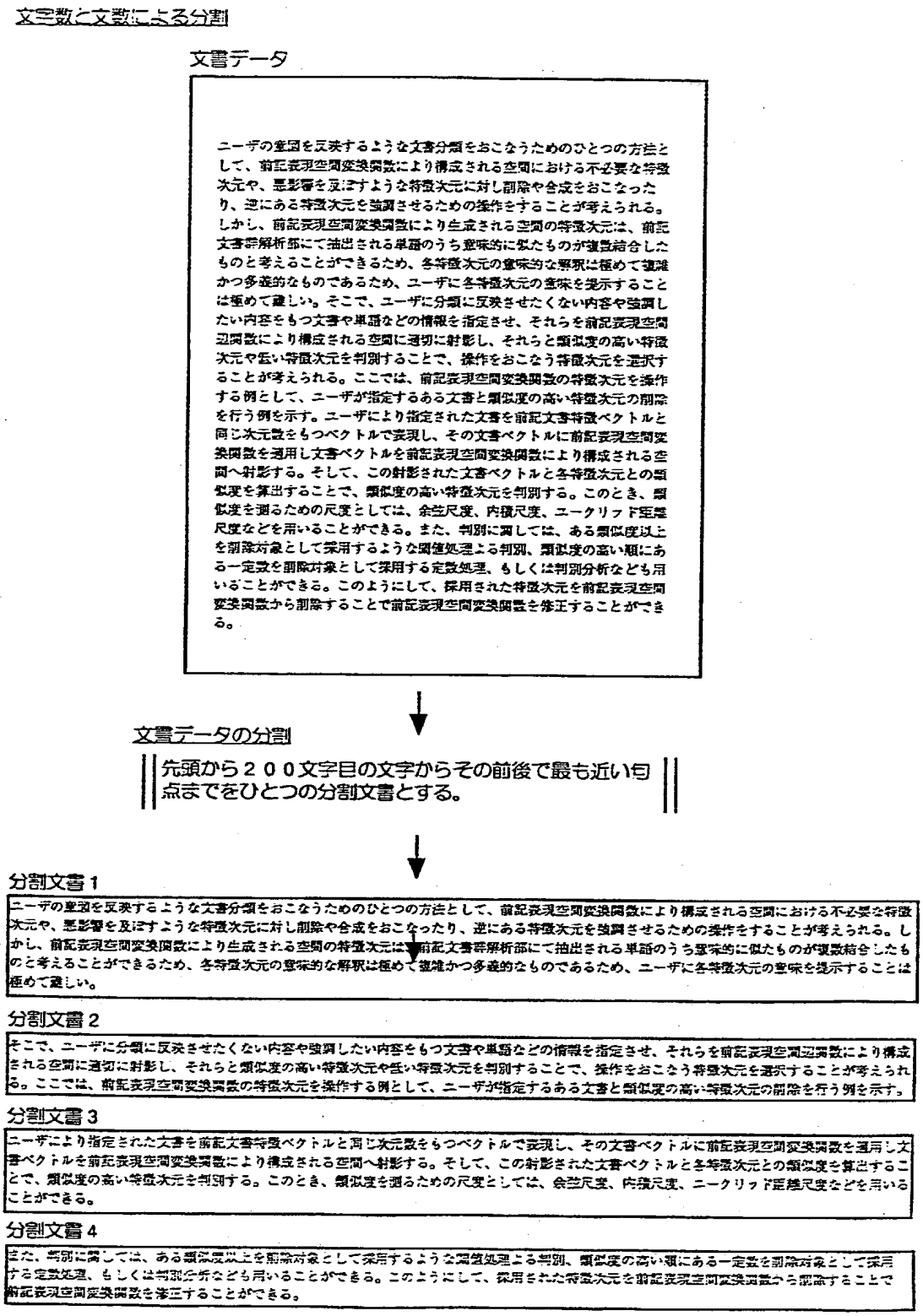
【図５２】


文書データ
"トピック"＋数字＋改行コード
ニューストピック（98/10/35）
トピック1
京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
トピック2
防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
トピック3
神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
トピック4
フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
トピック5
パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。
文書終端記号
自然言語処理
名詞（"トピック"と数字）と改行コードを抽出
名詞（"トピック"と数字）と改行コードの文書内での位置を抽出
"トピック"＋数字＋改行コードという文字列の位置を検出
文書データの分割
"トピック"＋数字＋改行コードという文字列を先頭とし、"トピック"＋数字＋改行コードという文字列、もしくは文書終端記号を終端として囲まれる文字列をひとつの分割文書データとする
分割文書データ
分割文書1
・京大農学部研究室でカドミウム検出の玄米茶を飲み学生が下痢などの症状。
分割文書2
・防衛庁調達実施本部証拠隠滅疑惑で同本部職員の親族宅にも資料を預けた。
分割文書3
・神奈川県はリストラ策一環の横浜消費生活センターを今年度で廃止表明。
分割文書4
・フィリピン航空が労使紛争の末、経営難となり57年の歴史に幕を引き廃業。
分割文書5
・パキスタンシャリフ首相が国連総会で来年9月迄にCTBT加盟の用意と表明。

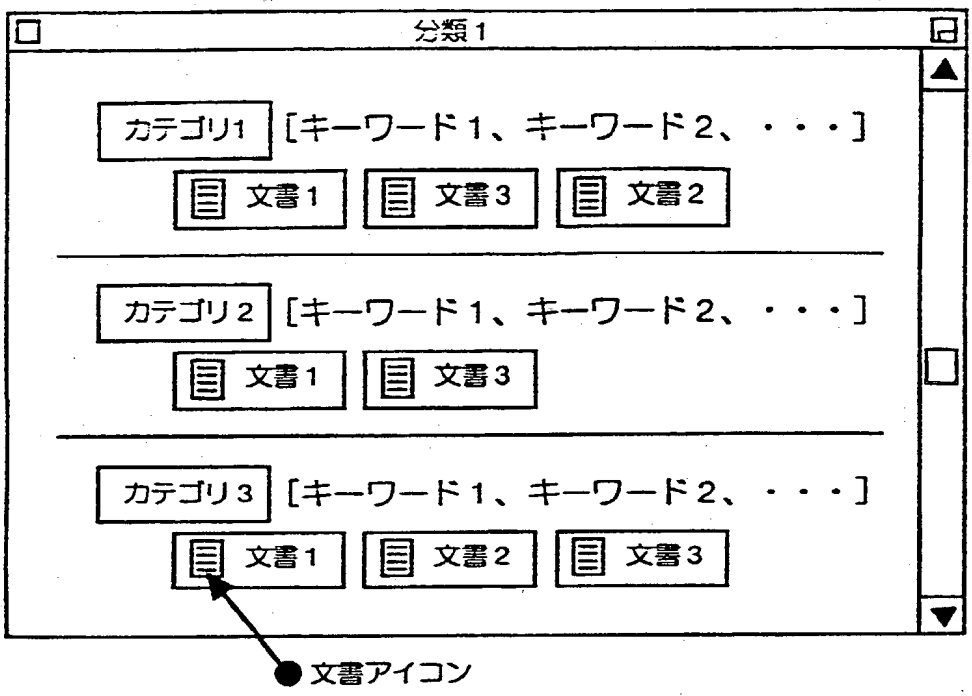
【図５３】


利用者による文書の分割
領域決定ライン
指示ポイント
分割文書とする領域
領域が決定している
ことを示す
分割文書としない領域
未決定領域
領域が決定していな
いことを示す
指示ポイントをドラッグす
ることで、上下に移動

【図５４】

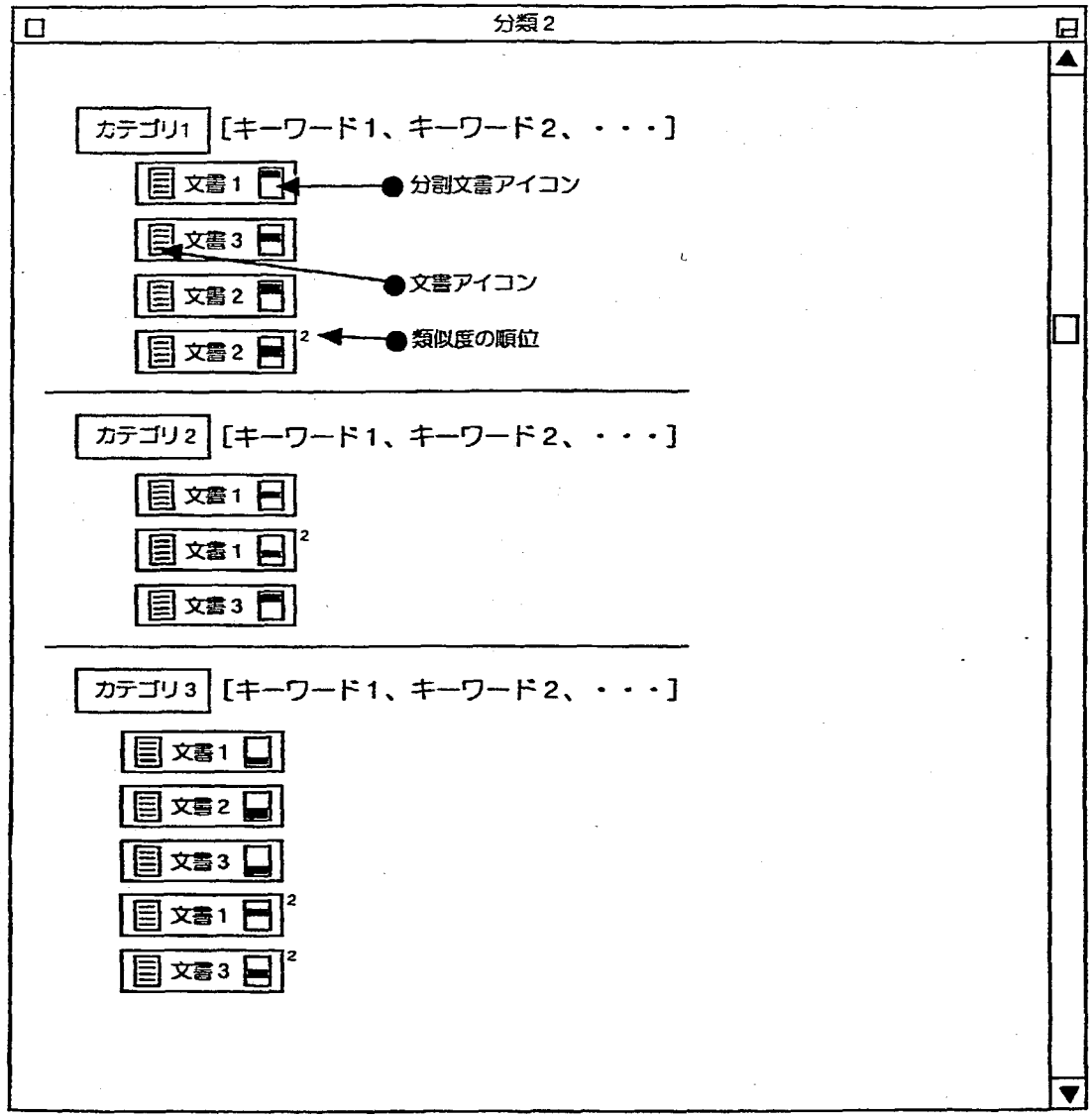

文字数と文数による分割
文書データ
ユーザの意図を反映するような文書分類をおこなうためのひとつの方法として、前記表現空間変換関数により構成される空間における不必要な特徴次元や、悪影響を及ぼすような特徴次元に対し削除や合成をおこなったり、逆にある特徴次元を強調させるための操作をすることが考えられる。しかし、前記表現空間変換関数により生成される空間の特徴次元は、前記文書解析部にて抽出される単語のうち意味的に似たものが複数結合したものと考えることができるため、各特徴次元の意味的な解釈は極めて複雑かつ多義的なものであるため、ユーザに各特徴次元の意味を提示することは極めて難しい。そこで、ユーザに分類に反映させたくない内容や強調したい内容をもつ文書や単語などの情報を指定させ、それらを前記表現空間辺関数により構成される空間に適切に射影し、それらと類似度の高い特徴次元や低い特徴次元を判別することで、操作をおこなう特徴次元を選択することが考えられる。ここでは、前記表現空間変換関数の特徴次元を操作する例として、ユーザが指定するある文書と類似度の高い特徴次元の削除を行う例を示す。ユーザにより指定された文書を前記文書特徴ベクトルと同じ次元数をもつベクトルで表現し、その文書ベクトルに前記表現空間変換関数を適用し文書ベクトルを前記表現空間変換関数により構成される空間へ射影する。そして、この射影された文書ベクトルと各特徴次元との類似度を算出することで、類似度の高い特徴次元を判別する。このとき、類似度を測るための尺度としては、余弦尺度、内積尺度、ユークリッド距離尺度などを用いることができる。また、判別に関しては、ある類似度以上を削除対象として採用するような閾値処理よる判別、類似度の高い順にある一定数を削除対象として採用する定数処理、もしくは判別分析なども用いることができる。このようにして、採用された特徴次元を前記表現空間変換関数から削除することで前記表現空間変換関数を修正することができる。
文書データの分割
先頭から２００文字目の文字からその前後で最も近い句点までをひとつの分割文書とする。
分割文書１
ユーザの意図を反映するような文書分類をおこなうためのひとつの方法として、前記表現空間変換関数により構成される空間における不必要な特徴次元や、悪影響を及ぼすような特徴次元に対し削除や合成をおこなったり、逆にある特徴次元を強調させるための操作をすることが考えられる。しかし、前記表現空間変換関数により生成される空間の特徴次元は、前記文書解析部にて抽出される単語のうち意味的に似たものが複数結合したものと考えることができるため、各特徴次元の意味的な解釈は極めて複雑かつ多義的なものであるため、ユーザに各特徴次元の意味を提示することは極めて難しい。
分割文書２
そこで、ユーザに分類に反映させたくない内容や強調したい内容をもつ文書や単語などの情報を指定させ、それらを前記表現空間辺関数により構成される空間に適切に射影し、それらと類似度の高い特徴次元や低い特徴次元を判別することで、操作をおこなう特徴次元を選択することが考えられる。ここでは、前記表現空間変換関数の特徴次元を操作する例として、ユーザが指定するある文書と類似度の高い特徴次元の削除を行う例を示す。
分割文書３
ユーザにより指定された文書を前記文書特徴ベクトルと同じ次元数をもつベクトルで表現し、その文書ベクトルに前記表現空間変換関数を適用し文書ベクトルを前記表現空間変換関数により構成される空間へ射影する。そして、この射影された文書ベクトルと各特徴次元との類似度を算出することで、類似度の高い特徴次元を判別する。このとき、類似度を測るための尺度としては、余弦尺度、内積尺度、ユークリッド距離尺度などを用いることができる。
分割文書４
また、判別に関しては、ある類似度以上を削除対象として採用するような閾値処理よる判別、類似度の高い順にある一定数を削除対象として採用する定数処理、もしくは判別分析なども用いることができる。このようにして、採用された特徴次元を前記表現空間変換関数から削除することで前記表現空間変換関数を修正することができる。

【図５５】



出証特平１１－３０９１７２６

【図５６】


分類２
カテゴリ1 [キーワード１、キーワード２、・・・]
文書１
分割文書アイコン
文書３
文書２
文書アイコン
文書２ 2
類似度の順位
カテゴリ２ [キーワード１、キーワード２、・・・]
文書１
文書１ 2
文書３
カテゴリ３ [キーワード１、キーワード２、・・・]
文書１
文書２
文書３
文書１ 2
文書３ 2

【書類名】　　要約書

【要約】

【課題】　文書の意味に係わるような分析作業において、単にその結果のみを出力するのではなく、情報分析作業全般にわたる支援をおこなうことを課題とする。

【解決手段】　入力された文書データを記憶する文書記憶部４０２と、文書記憶部４０２により記憶された文書データの全部または一部を選択する選択部４０３と、選択部４０３により選択された文書データの全部または一部から文字列の特徴に関するデータを抽出する特徴抽出部４０４と、特徴抽出４０４により抽出された文字列の特徴に関するデータに基づいて文書データの全部または一部を加工処理する加工処理部４０５と、加工処理部４０５により加工処理された文書データの全部または一部を出力する出力部４０６とを備える。

【選択図】　　図４

出願人履歴情報

識別番号　〔０００００６７４７〕

１．変更年月日　１９９０年　８月２４日
［変更理由］　新規登録
住　所　東京都大田区中馬込１丁目３番６号
氏　名　株式会社リコー